Probabilistic segmentation modelling and deep learning-based lung cancer screening

Benoît Audelan

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Modèles probabilistes pour la segmentation d'images et dépistage du cancer du poumon par apprentissage profond

Benoît Audelan

Inria, Équipe Epione

Thèse dirigée par Hervé Delingette

Soutenue le 22 juillet 2021

Présentée en vue de l'obtention du grade de Docteur en Automatique, Traitement du Signal et des Images de l'Université Côte d'Azur.

Devant le jury composé de :

Miguel A. González Ballester	Universitat Pompeu Fabra	Rapporteur
Bjoern H. Menze	Technische Universität München	Rapporteur
William M. Wells	Harvard Medical School	Examinateur
Charles Bouveyron	Université Côte d'Azur	Président
Hervé Delingette	Inria Sophia Antipolis	Directeur de thèse

UCA JEDI INITIATIVE D'EXCELLENCE

Inria

Epione e-patient / e-medicine

Modèles probabilistes pour la segmentation d'images et dépistage du cancer du poumon par apprentissage profond

Probabilistic segmentation modelling and deep learning-based lung cancer screening

Jury

Président du jury

Charles Bouveyron	Professeur	Université Côte d'Azur

Rapporteurs

Miguel A. González Ballester	Professeur	Universitat Pompeu Fabra
Bjoern H. Menze	Professeur	Technische Universität München

Examinateurs

William M. Wells	Professeur de radiologie	Harvard Medical School
Hervé Delingette	Directeur de recherche	Inria Sophia Antipolis

Probabilistic segmentation modelling and deep learning-based lung cancer screening

Abstract: This thesis is structured around two research themes dedicated to probabilistic image segmentation and lung cancer screening.

First, we focus on the problem of controlling the spatial regularity of segmentations. Enforcing a certain extent of regularization is important in order to guarantee the spatial consistency of the segmented structures and to control the smoothness of the segmentation boundaries. We investigate several probability distributions allowing spatial regularization to be enforced a priori in Bayesian segmentation models. Theses priors are incorporated into a common variational inference framework and compared with respect to several criteria, including their regularization strength, the complexity of their inference, their local adaptivity or their impact on uncertainty quantification.

In a second step, we address the challenge of controlling the quality of image segmentations in large databases when no reference segmentation is available. We propose a novel approach for unsupervised quality control based on a probabilistic model built on simple smoothness and intensity assumptions. Our method allows suspicious cases to be extracted from segmentation datasets and produces interpretable outputs enabling potential errors to be localized in the image.

Third, we develop a new approach for the fusion of continuous segmentation maps. By allowing a consensus between several experts or algorithms to be estimated, our method represents a new solution to tackle the problem of inter-rater variability. Our method, based on heavy-tailed distributions, allows for local variations in the raters' performances thus leading to a more robust consensus estimate. In addition, the concept of mixture of consensuses is introduced and its application to the clustering of raters is investigated.

The second focus of this thesis is dedicated to lung cancer screening from computed tomography images. We propose a fully automated pipeline based on deep learning. We provide an extensive analysis of the results of the pipeline on several datasets. In particular, we study the impact of training with subjective radiological labels, i.e. without any histopathologic ground truth, on the performances when applying the pipeline on real-life screening data.

Keywords: medical imaging, image segmentation, artificial intelligence, machine learning, deep learning, lung cancer.

Modèles probabilistes pour la segmentation d'images et dépistage du cancer du poumon par apprentissage profond

Résumé: Cette thèse s'articule autour de deux axes de recherche consacrés à la modélisation probabiliste de la segmentation d'images et au dépistage du cancer du poumon.

Dans un premier temps, nous nous intéressons au problème du contrôle de la régularité spatiale des segmentations d'images. Imposer un certain niveau de régularisation est important afin de garantir l'homogénéité des structures segmentées et de contrôler le lissage de leurs contours. Nous étudions différentes distributions de probabilité permettant d'imposer une régularisation spatiale a priori dans des modèles de segmentation probabilistes. Ces distributions sont regroupées au sein d'un même schéma d'optimisation basé sur l'inférence variationnelle, et sont comparées entre elles par rapport à plusieurs critères dont leur capacité à régulariser, la complexité de leur inférence, leur adaptabilité locale ou encore leur impact sur la quantification de l'incertitude.

Dans un deuxième temps, nous abordons le défi que constitue le contrôle de la qualité des segmentations au sein de grandes bases de données lorsqu'aucune segmentation de référence n'est disponible. Nous proposons une nouvelle approche pour le contrôle non supervisé reposant sur un modèle probabiliste construit sur des hypothèses simples d'intensité et de régularité. Notre méthode permet d'identifier des cas suspects et génère des sorties interprétables permettant de localiser dans l'image les erreurs potentielles.

Dans un troisième temps, nous développons une nouvelle approche pour la fusion de cartes de segmentation continues. En permettant d'estimer un consensus entre plusieurs experts ou algorithmes, notre méthode représente un nouveau levier face au problème de la variabilité pouvant être observée entre annotateurs. Notre approche, basée sur des distributions de probabilité à queue lourde, tient compte des variations locales des performances des annotateurs, ce qui permet d'obtenir une estimation plus robuste du consensus. De plus, le concept de consensus multiples est introduit et son application au regroupement des évaluateurs est étudiée.

Le second axe de cette thèse est dédié au dépistage du cancer du poumon à partir d'images tomodensitométriques. Nous proposons une chaîne de traitement entièrement automatisée utilisant l'apprentissage profond. Une analyse complète des résultats de la chaîne de traitement est réalisée sur plusieurs ensembles de données. En particulier, nous étudions l'impact sur les performances d'un entraînement réalisé uniquement avec des annotations radiologiques, c'est-à-dire sans vérité histopathologique, lors de l'application de la chaîne sur des données réalistes de dépistage.

Mots clés : imagerie médicale, segmentation d'images, intelligence artificielle, apprentissage artificiel, apprentissage profond, cancer du poumon.

Acknowledgement

I would like to thank first my thesis supervisor, Hervé Delingette, for your guidance all along these last three and a half years. In particular, I am deeply grateful for all the knowledge you allowed me to gather along the way. I would also like to thank Nicholas Ayache for accepting me as a member of the Epione team and giving me the opportunity to work in an outstanding and stimulating environment.

I am very grateful to Bjoern Menze and Miguel Ballester for reviewing carefully this manuscript and sharing their valuable and constructive remarks. I thank Charles Bouveyron and William Wells for accepting to be members of my jury.

This thesis has been very enriching from a professional point of view, but also from a personal one. It has given me the opportunity to meet and work with passionate, inspiring and wonderful people. I would therefore like to thank all the members of the Epione team. Thanks to Yann and Gaëtan for sharing my office during these years. In particular, Yann, thank you for your great kindness and your patience when you tried to give me some insights about Riemannian geometry. Thanks to Zihao for your good mood and the good times we had in Inria or in China, I couldn't have had a better guide in Shenzhen! Many thanks also to all the members of the office opposite my own, Clément, Luigi and Jaume, and also to Santiago, for all the moments and the discussions we shared together. I also thank Raphaël Sivera, Julian Krebs, Pawel Mlynarski, Tania Bacoyannis, Florent Jousse, Nicolas Cedilnik, Shuman Jia, Nicolas Guigui, Bastien Manach, Marco Milanesio and many more.

Je tiens à terminer ces remerciements par une pensée pour ma famille qui fut mon soutien le plus précieux au cours de ces trois dernières années.

Un grand merci à mes parents, pour m'avoir soutenu depuis toujours et m'avoir permis d'en arriver là aujourd'hui. Merci d'avoir toujours été présents à mes côtés, pour votre écoute face à mes nombreux questionnements et votre accompagnement sans faille tout au long de ces années d'études. Vous avez toujours réussi à trouver les mots, même dans les moments les plus difficiles. Je suis extrêmement reconnaissant pour tout ce que vous m'avez apporté et que quelques mots ne sauraient décrire.

Merci également à mes grand-parents, pour avoir toujours été derrière moi, et pour tout ce que vous m'avez transmis et dont je mesure la valeur inestimable aujourd'hui.

Merci aussi à Sophie, pour toutes les cartes et les discussions philosophiques.

Enfin, merci à Lucie. Merci pour ton soutien sans faille et pour avoir toujours gardé le sourire malgré les nombreuses difficultés occasionnées par la distance. Tu as toujours été là et tu as toujours réussi à me comprendre, même dans les moments les plus compliqués. Merci pour tout le soleil que tu m'apportes !

Thank you!

Benoît

Contents

1 Introduction **1**
- 1.1 Context of the thesis . . . 1
- 1.2 Objectives of the thesis . . . 4
- 1.3 Thesis overview . . . 4
- 1.4 List of publications . . . 5

2 Background and preliminaries **7**
- 2.1 Brief introduction to variational inference . . . 7
- 2.2 Brief introduction to lung cancer . . . 10

3 Spatial priors for Bayesian image segmentation **15**
- 3.1 Introduction . . . 16
- 3.2 Probabilistic segmentation framework . . . 19
- 3.3 Spatial smoothness priors . . . 20
  - 3.3.1 MRF/CRF priors . . . 20
  - 3.3.2 TV/FDSP priors . . . 21
  - 3.3.3 GP prior . . . 22
  - 3.3.4 GLSP prior . . . 23
- 3.4 Model inference . . . 25
  - 3.4.1 Label posterior approximation . . . 26
  - 3.4.2 Appearance parameters . . . 26
  - 3.4.3 Regularization variables . . . 27
  - 3.4.4 An incremental and sparse algorithm for the GLSP prior . . . 31
- 3.5 Results . . . 32
  - 3.5.1 Implementation of the algorithm . . . 32
  - 3.5.2 Whole image vs narrow band evolution . . . 34
  - 3.5.3 Spatial priors comparison . . . 35
  - 3.5.4 Uncertainty quantification . . . 38
  - 3.5.5 GLSP incremental algorithm . . . 40
- 3.6 Conclusion . . . 42

4 Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model **45**
- 4.1 Introduction . . . 46
- 4.2 Unsupervised quality control workflow . . . 49
  - 4.2.1 Input segmentation . . . 49
  - 4.2.2 Probabilistic model . . . 50
  - 4.2.3 Detection of challenging cases . . . 50

4.2.4 Use case . . . 51
4.3 Method . . . 51
4.3.1 Mixtures of multivariate Student's t-distributions . . . 51
4.3.2 Spatial smoothness prior . . . 51
4.3.3 Implementation . . . 53
4.4 Probabilistic inference . . . 55
4.4.1 MRF regularization . . . 55
4.4.2 GLSP regularization . . . 55
4.4.3 FDSP regularization . . . 55
4.5 Results . . . 57
4.5.1 Datasets . . . 57
4.5.2 Unsupervised indices . . . 58
4.5.3 Setting hyperparameters . . . 58
4.5.4 Qualitative analysis . . . 60
4.5.5 Quantitative analysis . . . 63
4.5.6 Results interpretability . . . 67
4.5.7 Surrogate segmentation performance . . . 69
4.5.8 Discussion . . . 70
4.6 Conclusion . . . 70

5 Robust Bayesian fusion of continuous segmentation maps 73
5.1 Introduction . . . 74
5.2 Robust estimate of consensus probability maps . . . 77
5.2.1 Baseline probabilistic framework . . . 77
5.2.2 Heavy-tailed distributions and scale mixture representation . . . 79
5.2.3 Model inference . . . 81
5.3 Mixture of consensuses . . . 85
5.3.1 Probabilistic framework . . . 85
5.3.2 Model inference . . . 85
5.4 Results . . . 87
5.4.1 Material . . . 87
5.4.2 Robust probabilistic framework . . . 89
5.4.3 Mixture of consensuses . . . 98
5.5 Discussion . . . 101
5.6 Conclusion . . . 102

6 End-to-end analysis of a computerized lung cancer screening pipeline based on LDCT 105
6.1 Introduction . . . 106
6.2 Material . . . 109
6.2.1 Training and validation: the LIDC-IDRI dataset . . . 109
6.2.2 Independent test sets . . . 112
6.3 Lung cancer screening pipeline . . . 113
6.3.1 Framework overview . . . 113
6.3.2 Lung segmentation network . . . 115
6.3.3 Nodule detection network . . . 116
6.3.4 Characterization network . . . 119
6.4 Results . . . 121

6.4.1 Cross-validation results on the LIDC-IDRI dataset . . . 121
6.4.2 Tests on independent datasets . . . 126
6.5 Discussion . . . 128
6.6 Conclusion . . . 132

7 Conclusion **135**
7.1 Main contributions . . . 135
7.2 Perspectives . . . 136
7.2.1 Spatial regularization in neural networks . . . 136
7.2.2 Unsupervised quality control of bounding boxes . . . 137
7.2.3 Consensus estimation in a supervised learning setting . . . 139
7.2.4 Towards an integrative analysis of chest CT scans . . . 141

Appendix A Spatial priors for Bayesian image segmentation **143**
A.1 Variational updates . . . 143
A.1.1 Update of $q(Z)$ . . . 143
A.1.2 Update of $q(\mathbf{W})$ . . . 144
A.1.3 Update of α . . . 146
A.1.4 Update of ξ and u . . . 146
A.2 Lower bound . . . 146
A.2.1 Expectations involving appearance parameters . . . 147
A.2.2 Expectations involving the label variable . . . 148
A.2.3 Expectations involving the spatial smoothness variables . . . 149
A.3 RKHS regularizers for the GLSP prior . . . 149
A.4 Derivation of the incremental algorithm for the GLSP prior . . . 151

Appendix B Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model **153**
B.1 Unsupervised indices . . . 153
B.1.1 Zeb [Zhang et al., 2008] . . . 153
B.1.2 F_{RC} [Zhang et al., 2008] . . . 154
B.1.3 η [Zhang et al., 2008] . . . 155
B.1.4 GS [Johnson & Xie, 2011] . . . 155
B.2 FDSP prior - variational inference . . . 155
B.2.1 Update of $q_Z^*(Z)$ . . . 156
B.2.2 Update of $q_{\mathbf{W}}^*(\mathbf{W})$ . . . 156
B.2.3 Update of $q_\alpha^*(\alpha)$ . . . 157
B.2.4 Update of $q_{\xi_n}^*(\xi_n)$ . . . 157
B.3 FDSP prior - lower bound . . . 157

Appendix C Robust Bayesian fusion of continuous segmentation maps **159**
C.1 Variational updates . . . 159
C.1.1 Robust probabilistic framework . . . 159
C.1.2 Mixture of consensuses . . . 163
C.2 Lower bound . . . 164
C.2.1 Robust probabilistic framework . . . 164
C.2.2 Mixture of consensuses . . . 166
C.3 Additional expectations . . . 166

C.3.1 Robust probabilistic model 167
C.3.2 Mixture of consensuses 167

Appendix D End-to-end analysis of a computerized lung cancer screening pipeline based on LDCT **169**
D.1 Circulant miRNAs as potential biomarkers for lung cancer detection . . 169
D.1.1 Introduction 169
D.1.2 Data collection and pre-processing 170
D.1.3 Experiments on the COSMOS dataset 171
D.1.4 Conclusion 173

Bibliography **175**

Chapter 1

Introduction

Contents

1.1 Context of the thesis 1
1.2 Objectives of the thesis 4
1.3 Thesis overview 4
1.4 List of publications 5

1.1 Context of the thesis

In this section, we introduce the motivations of the thesis structured around two main research themes. The first part is dedicated to image segmentation while the second focuses on lung cancer screening.

Image segmentation consists in assigning labels to pixels and results in a meaningful partition of the image into several homogeneous regions. It is a major image processing task, with applications in various domains, such as robotics and autonomous vehicles for object recognition and visual scene understanding [Cordts et al., 2016], agriculture for fertilization planning [Wang et al., 2013] but also, importantly, medical imaging. For instance, image segmentation plays a significant role in oncology, in particular in radiotherapy whose objective is the destruction of malignant cells by irradiation while preserving the normal tissues. Its planning involves the delineation of the gross tumor volume, which is a critical image segmentation step as it directly influences the extent of the radiation doses that will be administered to the patient [Mazzara et al., 2004]. Image segmentation is also used for screening purposes and its relevance for the diagnosis of COVID-19 has for example been demonstrated [Shi et al., 2021]. Computer-aided surgery relies also heavily on image segmentation for the detection and tracking of instruments in operation [Bodenstedt et al., 2018]. More generally, image segmentation is the cornerstone of many models or tools developed for patient monitoring and prognosis prediction [Sun et al., 2019; Ferdinand Christ et al., 2017; Liu et al., 2010].

A wide range of approaches has been proposed for automatic image segmentation, such as deep learning approaches [Garcia-Garcia et al., 2017], deformable models [Cremers et al., 2007], atlas-based methods [Iglesias & Sabuncu, 2015] or methods based on a probabilistic modelling of the image segmentation problem. Probabilistic segmentation models have several attractive properties. They rely on a few simple assumptions, leading to explainable outputs. It is a desirable property as interpretability is considered as essen-

tial, especially in the medical domain. Moreover, despite their apparent simplicity, these models provide good results and their relevance for brain segmentation has for example been demonstrated [Greenspan et al., 2006]. They may not not require any annotated training set, and their probabilistic nature allows uncertainty to be estimated.

Assuming some level of spatial regularization in the segmentation is one of the basic hypotheses commonly used in Bayesian segmentation modelling. Spatial regularization is important in order to take into account the spatial correlations between pixels, to guarantee the consistency of the final segmentation mask, but also to control the smoothness of the segmentation boundaries. The automatic estimation of the regularization parameters, the scalability of the approach to large images, and its local adaptivity are open challenges for research, and are the main topics of the third chapter of this thesis.

Moreover, the last few years have seen the rise of deep learning that now achieves state of the art performances in many computer vision tasks, including image segmentation [Kamnitsas et al., 2018; Chen et al., 2020]. Convolutional neural networks (CNNs) are specifically tailored to handle image inputs and have demonstrated a tremendous ability to extract automatically relevant features for the task of interest [LeCun et al., 2015]. Their architecture relies on a stack of layers where a set of local and non-linear operations are performed, such as convolution, pooling or up-sampling. These operations involve a large number of parameters, which are optimized during the training phase of the algorithm, by stochastic gradient error back-propagation.

CNNs realize their best performances in the supervised setting, which requires the availability of large amounts of labelled data. In the non-medical domain, several breakthroughs in performances could not have been achieved without the development of large-scale databases, such as the PASCAL Visual Object Classes (PASCAL VOC) dataset [Everingham et al., 2010] or the Common Objects in COntext (Coco) dataset [Lin et al., 2014], containing 9993 and more than 200K images, respectively. Providing ground truth annotations for so many images is a real challenge, addressed by leveraging crowd-sourcing for the COCO dataset.

Challenges are even greater in medical imaging, considering the legal obstacles to data collection and sharing, and the high level of expertise required for adding the annotations. Creating annotations is in addition time-consuming, because of the significant size (often 3D) and complexity of medical images. Yet, the development of large and completely annotated datasets is recognized as essential to enable the exploitation of the full potential of deep learning-based approaches in medical imaging [Langlotz et al., 2019; Willemink et al., 2020].

With the development of large datasets, annotation quality control is an issue of rising importance. Noisy segmentation labels lead to a severe drop in the performances of algorithms [Heller et al., 2018], and may also impact their evaluation and comparison. More generally, image segmentation is often only the pre-processing step of complex pipelines, whose outputs depend critically on the outcome of the segmentation [Unkelbach et al., 2014]. The question of ground truth quality control is also relevant for datasets annotated by medical experts. Manual delineations are indeed known to suffer from potentially large inter-rater variability [Louie et al., 2010; Genovesi et al., 2011; Nakamura et al., 2008]. If discrepancies between raters might reflect the uncertainty around the structure boundaries, they might also be the consequence of human errors, due, for example, to differences in levels of experience [Jeanneret-Sozzi et al., 2006; Kristensen et al., 2017]. Moreover, the segmentation of 3D images is often performed in 2D with a slice by slice approach, which can lead to inconsistencies after assembling the whole

volume [Crowe et al., 2017]. Yet, given the size of the datasets, visual inspection cannot be considered for reviewing all segmentation labels. Therefore, there is a crucial need in the development of automatic tools for quality control of image segmentation. In particular, there is a lack in generic and unsupervised methods that could be deployed on a wide range of image types in a convenient manner.

Image segmentation quality control can be a first step to mitigate the inter-rater variability issue. Another approach is to estimate a consensus between raters, using an appropriate data fusion method. The main objective of data fusion is to lead to a more robust estimate of the segmentation, which can be, for example, critical in radiotherapy planning [Li et al., 2009]. The estimation of a consensus between raters is a solution to the inter-rater variability issue, and has been well studied in the binary setting, where segmentations are discrete-valued maps. The most well-known algorithm is perhaps the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm [Warfield et al., 2004], which builds a consensus as a weighted combination of the raters' inputs, depending on their respective performances. In contrast, the continuous setting has received less attention, despite the fact that the interest of fusing multiple segmentation outputs produced by different algorithms, typically continuous, has been demonstrated, notably for deep learning outputs [Menze et al., 2015].

Given this context, two objectives of this thesis, investigated in the fourth and fifth chapters, are to explore novel approaches for segmentation quality control and robust consensus estimation. In particular, we investigate methods based on a probabilistic modelling of the image segmentation problem, applying the methodology developed in the third chapter.

The second main research theme of the thesis is related to lung cancer screening. Lung cancer is the leading cause of death by cancer in the world, with 1.8 millions of deaths worldwide in 2020, according to the World Health Organization[1]. It is thus considered as a major public health problem, with economic and social consequences. Lung cancer can be defined as an uncontrolled growth of tumor cells that originate in lung tissue. As for other cancers, the stage of the disease at diagnosis is critical and directly influences the chance of survival of the patient. Yet, many lung cancers are currently detected at late stages, due to the late onset of symptoms. Moreover, the limited understanding of the tumor biology and the lack of effective treatments are also factors explaining the poor prognosis of lung cancer. According to the National Cancer Institute[2], the 5-year survival rate is 20.5% in the United States. There is therefore a need in screening strategies to enable early lung cancer detection.

Two large studies, the American National Lung Screening Trial (NLST) and the Dutch-Belgian NELSON trial have highlighted the potential of medical imaging, in particular low-dose computed tomography (LDCT), for lung cancer screening, with a reduction of 20% in mortality [NLST, 2011; Koning et al., 2020]. LDCT screening is a radiological task consisting in the examination of a 3D image in search of small pulmonary abnormalities, a.k.a. nodules, suspicious for lung cancer. Yet, a thoracic LDCT scan contains typically several millions of pixels, making the search of nodules one of the most difficult and time-consuming task for radiologists.

[1] https://gco.iarc.fr/today/data/factsheets/cancers/15-Lung-fact-sheet.pdf, accessed April 3, 2021.
[2] https://seer.cancer.gov/statfacts/html/lungb.html, accessed April 3, 2021.

With the implementation of LDCT screening at large scales, such as in the United States [Wood et al., 2018], there is a need in the development of automatic tools to assist the radiologist in the CT scan examination. In particular, the detection and the characterization of nodules are two challenging computer vision tasks. In the literature, several approaches have already been proposed, either for the detection of nodules or their classification. However, less results have been reported regarding the integration of the two tasks in fully automated pipelines.

Moreover, current state of the art methods are based on deep learning [Liao et al., 2019]. Yet, they are often trained and evaluated on datasets with uneven label quality. For instance, the subject cancer status or reliable nodule labels are sometimes missing, as in the widely used publicly available LIDC-IDRI dataset [Armato III et al., 2011]. Therefore, there is also a need in assessing the performance of these algorithms on independent datasets, where reliable annotations are provided. This issue constitutes the core of the study reported in the last chapter of this thesis.

1.2 Objectives of the thesis

We now summarize the main objectives of this thesis, given the context described above. The first is to explore approaches for enforcing spatial regularization inside probabilistic image segmentation models. The second is to propose novel ideas to unsupervised segmentation quality control and robust consensus estimation. Finally, the thesis aims at investigating the relevance of a fully automated lung screening pipeline based on LDCT. These objectives are associated to the following questions:

- How to control the regularity of a segmentation in a data-driven way? In particular, how to enforce spatial regularization within a Bayesian segmentation framework?
- How to assess the annotation quality of an image segmentation dataset, containing potentially a very large number of cases, without any prior knowledge about the segmented structures?
- How to fuse several continuous input maps into a consistent and robust consensus?
- What are the real performances of a deep learning-based lung screening pipeline, solely trained with subjective radiological labels?

1.3 Thesis overview

In **chapter 2**, we discuss some methodological tools used in the remainder of the thesis. In particular, we introduce the rationale of variational inference. We also present a general overview of lung cancer.

In **chapter 3**, we focus on a Bayesian formulation of the binary image segmentation problem, leading to a probabilistic output. Bayesian segmentation models are commonly based on the combination of appearance and spatial regularization terms. We study the properties of several probabilistic smoothness label priors. A comparison is performed according to several criteria, such as the complexity of their inference, their regularization strength, and their influence on the model uncertainty estimation. In particular, all priors are unified into a common inference framework relying on variational methods.

In **chapter 4**, we focus on the important problem of the quality assessment of image segmentations. We propose a novel automated approach, based on a generic probabilistic model, enabling to identify difficult, potentially suspicious, segmentation cases within a large dataset. In particular, our proposed approach is unsupervised, meaning that it is not specific to the structure being segmented and can be applied conveniently on any database with segmentation labels. Moreover, we show that the approach is highly interpretable and allows potential errors to be localized within the image, which is of high interest for the medical domain, given the large sizes of the medical images.

In **chapter 5**, we propose a novel approach to the fusion of probabilistic maps. We address several limitations of previous works in continuous consensus estimation. The novelty of our method lies in the introduction of heavy-tailed distributions, enabling the evaluation of the raters' performances to be spatially adaptive. Moreover, a novel approach for the clustering of raters is investigated, based on the original concept of mixture of consensuses.

In **chapter 6**, we analyse the performances of a fully automated lung screening pipeline based on LDCT. Our framework relies entirely on deep learning and is composed of 3 networks, corresponding to the 3 following steps: lung parenchyma segmentation, nodule detection, and nodule characterization. We investigate the impact of unreliable labels by training our pipeline on the publicly available LIDC-IDRI dataset and then by conducting an analysis of its performances on several independent test sets, including real life screening data with high quality annotations.

Finally, in **chapter 7**, we summarize the main contributions of the thesis. We conclude with a discussion of potential future work and perspectives.

1.4 List of publications

This thesis led to the following peer-reviewed publications:

Journal articles

- [Audelan & Delingette, 2021] **B. Audelan** and H. Delingette. Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model. *In Medical Image Analysis, 68, 2021, p. 101895.*

- Robust Bayesian fusion of continuous segmentation maps. *In preparation for submission to a journal.*

- [Heeke et al., 2019] S. Heeke, J. Benzaquen, E. Long-Mira, **B. Audelan**, et al. In-house Implementation of Tumor Mutational Burden Testing to Predict Durable Clinical Benefit in Non-small Cell Lung Cancer and Melanoma Patients. *In Cancers, 11.9, 2019.*

Conference papers

- [Audelan & Delingette, 2019] **B. Audelan** and H. Delingette. Unsupervised Quality Control of Image Segmentation Based on Bayesian Learning. *In Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pp. 21–29.*

- [Audelan et al., 2020] **B. Audelan**, D. Hamzaoui, S. Montagne, R. Renard-Penna and H. Delingette. Robust Fusion of Probability Maps. *In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 259–268.*

Conference abstracts

- [Audelan et al., 2021] **B. Audelan**, S. Lopez, P. Fillard, Y. Diascorn, B. Padovani and H. Delingette. Validation of lung nodule detection a year before diagnosis in NLST dataset based on a deep learning system. *Submitted to ERS International Congress 2021.*

Chapter 2

Background and preliminaries

Contents

2.1 Brief introduction to variational inference 7
2.2 Brief introduction to lung cancer 10

This introductory chapter develops important concepts relevant for the remaining of the thesis. In particular, the first section recalls the basics of variational inference, a framework enabling the tractability of probabilistic models that will be leveraged in the next 3 chapters of the thesis. The second section focuses on lung cancer and discusses the etiology, the diagnostic and the current treatment protocols.

2.1 Brief introduction to variational inference

As presented in the introduction, the first research theme of the thesis is related to image segmentation and, in particular, to probabilistic segmentation modelling. A Bayesian model classically involves observed variables, hidden variables, and some hyperparameters θ. In an image segmentation context, the observed variables, that will be referred to as I, represent image-derived information, typically the image intensity. In contrast, the values of the hidden variables, also called latent variables, denoted here as $\mathbf{W}$, are unknown.

The main objective is then to estimate the hidden variables according to the available data. In other terms, one wants to compute the posterior probability $p(\mathbf{W}|I,\theta)$ of the hidden variables given the observed ones. The relationship between variables is described by the well-known Bayes' theorem:

$$p(\mathbf{W}|I,\theta) = \frac{p(I|\mathbf{W})p(\mathbf{W}|\theta)}{p(I|\theta)}\,, \tag{2.1}$$

where $p(I|\mathbf{W},\theta)$ is the likelihood specifying how I relates to $\mathbf{W}$, $p(\mathbf{W}|\theta)$ is the prior and $p(I|\theta)$ is the model evidence. The prior is intended to reflect some initial knowledge that one might have about the model output. A wide range of priors have been developed for image segmentation, introducing for example constraints related to the shape of the segmented structure [Nosrati & Hamarneh, 2016]. In particular, the third chapter of this thesis is dedicated to the investigation of priors enforcing spatial regularization, thus controlling the smoothness of the final mask. The model evidence can be re-written as $\int_{\mathbf{W}} p(I|\mathbf{W})p(\mathbf{W}|\theta)d\mathbf{W}$, which explains why it is also called marginal likelihood. The

model evidence can indeed be seen as a likelihood function in which some latent variables have been marginalized.

In many cases, the main obstacle to the evaluation of the posterior is the intractability of the model evidence. In other words, the computation of the integral does not lead to a closed-form analytical solution and therefore requires to be approximated. Methods to approximate the posterior are generally divided in 2 categories, depending on whether they are based on deterministic or stochastic approaches [Bishop, 2006].

Markov Chain Monte Carlo (MCMC) is a typical example of a stochastic approach allowing the posterior to be approximated. It relies on an iterative sampling scheme that leads theoretically to the exact posterior density. In comparison, deterministic approaches are based on the optimization of an explicit objective, the marginal likelihood, and are more efficient in most cases [Salimans et al., 2015].

The Laplace method is the most simple but widely used deterministic approach, which performs a second-order Taylor expansion at the posterior mode leading to a Gaussian approximation. The approximation is valid locally around the mode, which can be found with numerical optimization techniques. The approach of alternating between a Laplace approximation step for the posterior and a maximization step of the marginal likelihood to estimate the model hyperparameters θ is known as the evidence framework or empirical Bayes, or also as the type-II maximum likelihood method [Bishop, 2006; MacKay, 1999].

Variational inference is another major deterministic method to approach the posterior. The rationale of variational inference arises when decomposing the log marginal likelihood as follows:

$$\log p(I|\theta) = \underbrace{\int_{\mathbf{W}} q(\mathbf{W}) \log \frac{p(I, \mathbf{W}|\theta)}{q(\mathbf{W})} d\mathbf{W}}_{\mathcal{L}(q)} + \underbrace{\int_{\mathbf{W}} q(\mathbf{W}) \log \frac{q(\mathbf{W})}{p(\mathbf{W}|I, \theta)} d\mathbf{W}}_{\mathrm{KL}[q(\mathbf{W})||p(\mathbf{W}|I,\theta)]}, \tag{2.2}$$

where $q(\mathbf{W})$ denotes any probability distribution defined over the latent variables. The second term in Eq. 2.2 is the Kullback-Leibler (KL) divergence between the distribution $q(\mathbf{W})$ and the posterior $p(\mathbf{W}|I, \theta)$. By definition of a KL divergence, we have $\mathrm{KL}[q(\mathbf{W})||p(\mathbf{W}|I, \theta)] \geq 0$ with equality if and only if the two distributions are equal. Thus, the best approximation $q^*(\mathbf{W})$ to the true posterior is obtained when the divergence vanishes. One can note that this is what happens during the maximization step of the expectation-maximization (EM) algorithm: the approximation by a Dirac distribution (point estimation) is updated by computing the true posterior. However, we consider here more general cases for which the true posterior is intractable and therefore for which the EM procedure is not appropriate.

The KL divergence being always positive, the first term in Eq. 2.2 is a lower bound over the log marginal likelihood and is therefore sometimes referred to as the Evidence Lower Bound (ELBO). As the log marginal likelihood $\log p(I|\theta)$ is not a function of $q(\mathbf{W})$, one can see that maximizing the ELBO with respect to $q(\mathbf{W})$ amounts to minimize the KL divergence. This is exactly the objective of variational inference: instead of trying to minimize directly the KL divergence, it maximizes the lower bound, which can be computed for well-chosen families of probability distributions and enables actually to solve the same problem [Bishop, 2006].

An alternative to variational inference is the expectation-propagation (EP) algorithm that seeks to minimize the reverse KL divergence, which is written $\mathrm{KL}[p(\mathbf{W}|I, \theta)||q(\mathbf{W})]$.

However, this approach can lead to a poor approximation if the true posterior is multimodal, as the EP method tries to capture all the modes of the distribution. In contrast, variational inference concentrates the probability mass of the approximation such that it matches the regions with high probability mass of the true posterior. This leads to better results for multimodal distributions, which are very common in practice.

The variational approach is based on the tractability of the evidence lower bound. This tractability can be guaranteed by introducing some assumptions regarding the form of the approximate distribution, $q(\mathbf{W})$, that was for now unconstrained. In particular, variational inference restricts the space of possible probability distributions to some chosen families that lead to closed-form solutions [Blei et al., 2017]. The objective is then to find the member of these families that approximates the best the true posterior $p(\mathbf{W}|I, \theta)$.

The mean field approximation is an example of hypothesis that can be made about the approximate posterior and that is widely used in practice. Consider a model with N latent variables, i.e. $\mathbf{W} = \{w_i\}_{1 \leq i \leq N}$. The mean field approximation assumes a factorization of the approximate posterior with respect to each hidden variable, such that:

$$q(\mathbf{W}) = \prod_{i=1}^{N} q_i(w_i)\,. \tag{2.3}$$

Note that this is the only assumption made about the approximate distribution. In particular, no further conditions are imposed on the form of each individual factor $q_i(w_i)$.

After the insertion of the factorized form in the lower bound, the analysis proceeds by examining one of the factor $q_j(w_j)$ while considering the others constant. Noting $\mathbf{W}_i = \{w_i\}_{i \neq j}$, the lower bound can be re-written as:

$$\begin{aligned}
\mathcal{L}(q) &= \int_{\mathbf{W}} q_j(w_j) \prod_{i \neq j} q_i(w_i) \Big[\log p(I, \mathbf{W}) - \log q_j(w_j) \Big] d\mathbf{W} + cst\,, \\
&= \int_{w_j} q_j(w_j) \Big[\Big(\int_{\mathbf{W}_i} \prod_{i \neq j} q_i(w_i) \log p(I, \mathbf{W}) d\mathbf{W}_i \Big) - \log q_j(w_j) \Big] dw_j + cst\,, \\
&= \int_{w_j} q_j(w_j) \log \frac{\exp \mathbb{E}_{q_{i \neq j}}[\log p(I, \mathbf{W})]}{q_j(w_j)} dw_j + cst\,, \\
&= -\mathrm{KL}\left(q_j(w_j) || \exp \mathbb{E}_{q_{i \neq j}}[\log p(I, \mathbf{W})] \right) + cst\,,
\end{aligned} \tag{2.4}$$

where $\mathbb{E}_{q_{i \neq j}}[\log p(I, \mathbf{W})]$ is the expectation of the model log joint probability with respect to the distribution $\prod_{i \neq j} q_i(w_i)$ containing $N-1$ factors. Because the divergence is always positive, the minimum of the lower bound with respect to $q_j(w_j)$ is achieved when it vanishes, leading to the following important result:

$$\log q_j^*(w_j) = \mathbb{E}_{q_{i \neq j}}[\log p(I, \mathbf{W})] + cst\,. \tag{2.5}$$

This analysis is valid for any factors and leads to an iterative algorithm in which the posterior approximations of the latent variables are optimized in turn, while considering the others constant. Convergence to a local optimum is guaranteed because of the convexity of the lower bound with respect to each of the factors. Each iteration

corresponds to an increase in the lower bound, which can be computed to assess the convergence and to perform model selection [Blei et al., 2017].

2.2 Brief introduction to lung cancer

The human body has 2 lungs, a right lung and a left lung, located on either side of the mediastinum inside the thoracic cavity, as shown in Fig. 2.1. The right and left lungs are further divided into 3 and 2 territories, called lobes. The lungs are part of the lower respiratory tract that begins with the trachea. After the bronchi, the airways ramify into smaller and smaller ducts up to the alveoli, where the gas exchange takes place. The lungs are isolated from the thoracic cage by the pleura, a membrane enabling to reduce the frictions caused by the respiratory motion.

The lung tissues are a complex community of cells playing a vital role in the respiration. Yet, this environment is fragile and sensitive to toxic agents. An accumulation of alterations over the years can lead to the onset of an uncontrolled growth of tumor cells, and thus to cancer. Lung cancer is a major health problem in the world, representing 11.4% of the new cancer cases and 18% of the deaths by cancer in 2020, according to World Health Organization[1]. The highest incidence rates are found in North America, Europe and eastern Asia, but data are difficult to compare due to the lack of reporting in some countries.

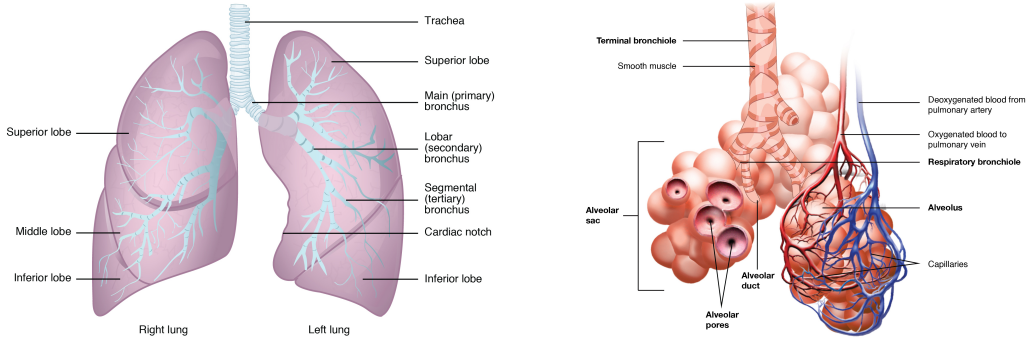


Fig. 2.1: Representation of the general anatomy of the lungs and airways (2.1a). Focus on the respiratory zone (2.1b). (Figures from [Betts et al., 2013], CC-By license.)

Lung cancer incidence is relatively low before 40 years old, then increases slowly to peak between 65 and 84 years old. In the United States, the median age at diagnosis is 71 years old [Duma et al., 2019]. The advanced age of the population at risk is often associated with multiple comorbidities, which makes the management of the patient more difficult.

The primary risk factor for lung cancer is active or passive smoking: it is estimated that 80% of men and 50% of women diagnosed with lung cancer have a smoking history [Gridelli et al., 2015]. Nonetheless, around 10% of lung cancer patients in the United States are never smokers [Couraud et al., 2012]. Exposure to air pollution and toxic substances such as radon, asbestos, cooking fumes, is known to be an increased risk factor.

[1] https://gco.iarc.fr/today/data/factsheets/cancers/15-Lung-fact-sheet.pdf, accessed April 3, 2021

More generally, family history and living conditions may influence the development of lung cancer.

A majority of lung cancer cases are diagnosed in symptomatic patients. Common symptoms are cough, chest pain, dyspnea, weight loss and hemoptysis. Exact diagnosis requires imaging and tissue sampling for pathological analysis.

Computed tomography (CT) is currently the imaging gold standard for lung cancer diagnosis [Tsim et al., 2010]. This method is based on measuring the absorption of X-rays in the tissues, which varies according to the type of tissue. Absorption levels are recorded from different angles and then processed by algorithms to reconstruct a 3D image. The voxel intensity reflects the degree of X-rays absorption of the corresponding volume in the body and is measured in Hounsfield units. The Hounsfield scale ranges from -1000 HU for the air to $+1000$ HU for dense bone, with 0 HU being the attenuation value of water. CT scans offer a high spatial resolution, enabling a precise estimation of the tumor size, a search for any mediastinal or vascular invasion, or for lymph node involvement [Tsim et al., 2010]. Moreover, it provides valuable anatomical information for planning an eventual surgical resection.

Other imaging techniques include the positron emission tomography (PET) and the magnetic resonance imaging (MRI). The former uses a radioactive tracer to detect regions of abnormally high metabolic activity, typical of cancer cells. Its spatial resolution is however significantly lower than CT. A combination of the two (PET-CT) can be considered to obtain both metabolic and morphological data [Kaseda, 2020]. MRI is based on the measurement of signals related to the nuclear magnetic resonance of atoms in the body. Its application to lung cancer detection is limited because of its sensitivity to the respiratory motion and the low proton density of lung parenchyma leading to a lower signal-to-noise ratio [Kumar et al., 2016]. However, MRI is recommended in lung cancer patients for the detection of brain metastases.

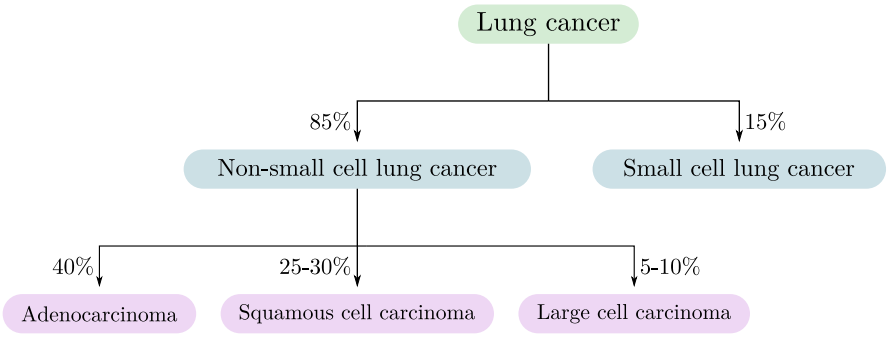


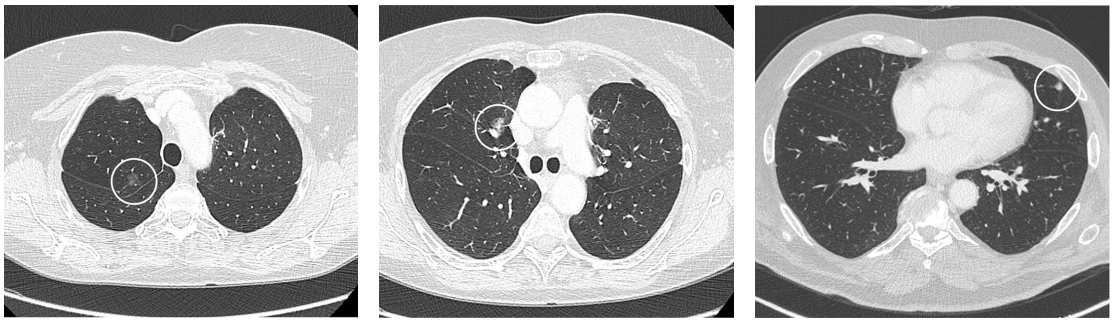
Fig. 2.2: Lung cancer classification.

Imaging and tissue samples are essential for a precise characterization of the tumor and staging of the disease. Lung cancer presents indeed a high level of heterogeneity with several histologic subtypes detailed in Fig. 2.2. The classification is based on the histologic origin of the tumor cells. Non-small cell lung cancer (NSCLC) is the most common type of lung cancer and can itself be divided into 3 subtypes [Zheng, 2016]. Adenocarcinoma originates in cells of the smaller airway epithelium with a glandular differentiation. Squamous cell carcinoma arises from cells with a squamous differentiation located in the central lung or in the major airways. The third subtype, large cell carcinoma, is used to regroup poorly differentiated tumors.

The staging of lung cancer is based on the tumor node metastasis (TNM) system, which provides a score reflecting the extent of the primary tumor and of metastases in other parts of the body [Tsim et al., 2010]. It relies on 3 criteria, namely the characteristics of the primary tumor, the involvement of lymph nodes and the presence of metastases. Accurate staging is important, as it impacts the treatment options and the management of the patient.

Early stage patient without any contraindication can be offered surgical resection. However, the surgical approach remains highly complex and patient dependent. Other possibilities include radiotherapy and chemotherapy, the latter being particularly suitable to target distant metastases sites. Treatment of later stages is more complex and has for a long time been associated with a poor prognosis. In the last few years, targeted therapy and immunotherapy have revolutionized the treatment of cancers and offered new perspectives [Duma et al., 2019]. The former relies on drugs that target proteins playing a role in cancer development in order to inhibit their activity. The latter aims at re-activating the immune defense against malignant cells. However, the effect of targeted therapy is often limited by the development of drug resistance over time. Moreover, immunotherapy may lead to serious adverse effects and benefits durably to only a small minority of patients. The mechanisms behind these limitations are not fully understood and are an active topic of research.

Despite recent advances in treatment, the prognosis for late stages patients remains low with a 5-year survival rate of 5% in the United States[2]. Thus, being able to detect lung cancer early before the onset of symptoms is critical.

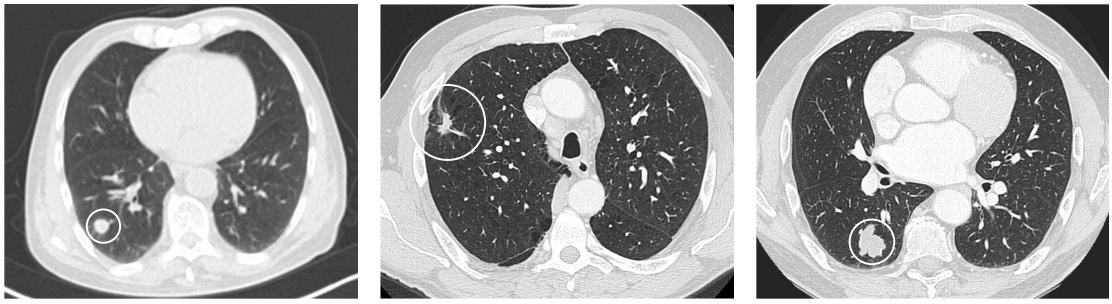


(a) **(b)** **(c)**

Fig. 2.3: Illustration of the variety of nodule appearances: example of a ground-glass nodule (2.3a), of a part-solid nodule (2.3b) and of a solid nodule (2.3c). (Figures adapted from [Snoeckx et al., 2018], CC-By license.)

Low-dose computed tomography (LDCT) has been demonstrated to be effective for lung cancer screening and is now recommended by the United States Preventive Services Task Force (USPSTF) for subjects at risk [USPSTF, 2021]. Radiation doses in LDCT are reduced in comparison with conventional CT to limit the exposure while maintaining an acceptable image quality [Rampinelli et al., 2013]. Lung cancer screening by LDCT involves to look for nodules and assess their malignancy. The detection of lesions is a first challenge: nodules are lung opacities up to 3 cm in diameter, thus very small compared to the whole image. Assessing their malignancy from the image is even more difficult: many factors other than cancer can lead to the apparition of nodules

[2] https://seer.cancer.gov/statfacts/html/lungb.html, accessed April 3, 2021.



(a) (b) (c)

Fig. 2.4: Illustration of the variety of nodule margins: example of a smooth margin (2.4a), of a spiculated margin (2.4b) and of a lobulated margin (2.4c). (Figures adapted from [Snoeckx et al., 2018], CC-By license.)

and the majority are benign. Nevertheless, several features have been reported as being associated with increased likelihood of malignancy [Snoeckx et al., 2018; MacMahon et al., 2005; Erasmus et al., 2000]. For example, nodules are commonly divided into 3 classes depending on their attenuation values, as shown in Fig. 2.3, and part-solid nodules are more suspicious for lung cancer.

Moreover, large diameters are also recognized as an increased risk factor. The diversity of the nodule margins, illustrated in Fig. 2.4, may also be taken into account: nodules presenting patterns of lobulation or spiculation are more likely to be malignant. However, the main limitation of these features is their limited specificity: two nodules may share the same morphological characteristics despite opposite malignancy status. LDCT screening remains therefore one of the most challenging radiological task: the detection of lesions is tedious and time-consuming, while the malignancy assessment is complex and subjective, which requires years of experience.

Chapter 3

Spatial priors for Bayesian image segmentation

Contents

3.1 Introduction **16**
3.2 Probabilistic segmentation framework **19**
3.3 Spatial smoothness priors **20**
3.3.1 MRF/CRF priors 20
3.3.2 TV/FDSP priors 21
3.3.3 GP prior 22
3.3.4 GLSP prior 23
3.4 Model inference **25**
3.4.1 Label posterior approximation 26
3.4.2 Appearance parameters 26
3.4.3 Regularization variables 27
3.4.4 An incremental and sparse algorithm for the GLSP prior . . 31
3.5 Results **32**
3.5.1 Implementation of the algorithm 32
3.5.2 Whole image vs narrow band evolution 34
3.5.3 Spatial priors comparison 35
3.5.4 Uncertainty quantification 38
3.5.5 GLSP incremental algorithm 40
3.6 Conclusion **42**

Image segmentation is a key image processing task resulting in the partition of an image into multiple regions. It is especially important in the medical domain as a starting point for many clinical downstream applications. We focus in this chapter on a Bayesian formulation of the binary image segmentation problem that leads to a probabilistic output. The Bayesian setting combines image-driven information with spatial regularization terms, because spatial consistency is generally considered as a desirable property for image labels. The objective of the chapter is to study the properties of several probabilistic smoothness label priors defined on a discrete or continuous domain. We also introduce a novel label prior based on Gaussian processes, whose inference is made tractable due to the periodic or separable nature of the covariance matrices. We compare these priors

with respect to several criteria including the complexity of their inference, their ability to spatially regularize image labels, and their impact on uncertainty quantification. We propose a common tractable inference scheme based on variational Bayes methods and the maximization of well-chosen local lower bounds over the log likelihood. We show how this generic framework can be used for post-processing regularization after any image segmentation algorithm but also as a standalone segmentation method. In particular, we introduce a novel segmentation strategy, based on the evolution of an isoprobability surface inside a narrow band.

3.1 Introduction

Image segmentation is one of the most studied problems in computer vision and a large range of frameworks has been proposed to handle this task [Nosrati & Hamarneh, 2016].

Building a proper image partition requires taking into account the relationships between pixels. Indeed, the output of a segmentation algorithm is expected to present some level of spatial homogeneity reflecting the consistency of the real object. In practice, the regularization strategy can vary depending on the segmentation framework. For instance, the level-set approach [Chan & Vese, 1999] introduces spatial smoothness by penalizing the segmentation area and the length of the segmentation boundaries. More recently, deep learning-based methods have met with considerable success, mainly in supervised image segmentation [Litjens et al., 2017]. Classical losses such as cross entropy or soft Dice do not explicitly constraint spatial consistency. The model is expected to learn the correct level of spatial regularization from the training database and by involving convolution kernels in the neural networks.

Another issue of increasing importance in image segmentation is the quantification of the uncertainty in the estimation of segmented structure boundaries. This is especially true in the medical domain as segmentation results impact decisions that may be made about the patient. For instance, in radiotherapy planning, the delineation of tumor lesions directly influences the extent of the dose delivered to the patient. Also, the RECIST criteria in oncology that guide the monitoring of patients with detected tumor lesions are based on the evolution of their estimated volumes. In this case, a realistic uncertainty estimate can be obtained by generating plausible segmentation samples and then computing a distribution of their volumes. Approaches relying on level sets or graphs [Boykov & Jolly, 2001; Rother et al., 2004] only give the most probable segmentation and therefore do not allow new samples to be generated. deep learning-based methods have achieved considerable success in supervised and unsupervised image segmentation [Litjens et al., 2017] and Monte Carlo dropout has been proposed to perform stochastic sampling through random perturbations of the weights in the network [Gal & Ghahramani, 2016; Nair et al., 2020]. Yet, a proper assessment of the uncertainty in the results remains an open question [Jungo & Reyes, 2019].

In this chapter, we focus on probabilistic approaches of the image segmentation problem. Those methods allow for a data-driven estimation of their parameters, as is the case for deep learning algorithms. However, their restricted number of parameters make them able to process images with little or no supervision, thus avoiding the need for annotating large image databases. In generative probabilistic models, the parameters can be fully interpreted which is a major advantage for detecting any algorithm failure. Another interesting feature of probabilistic models is that the Bayesian setting allows

for a proper uncertainty assessment by estimating the posterior probability of its output. In addition, the estimation of marginal likelihood enables model selection, i.e., selecting the model which best suits the data. Yet, the inference of those probabilistic models for image segmentation is relatively slow compared to the execution of a trained neural network. Furthermore, the definition of hand-crafted imaging features in probabilistic models is complex and often less discriminative than a supervised convolutional neural network (CNN). This is why there is active research towards the development of Bayesian neural networks [Wang & Yeung, 2016] to limit the overfit typical of classical CNNs, to train networks with less annotated data, to decrease the size of networks or to have a better estimation of their uncertainty.

Probabilistic models for image segmentation classically include a first component describing the appearance of the region of interest and a second constraining its shape. A wide range of priors has been proposed for image segmentation, relying for instance on knowledge of the appearance or the shape of the segmented structure [Nosrati & Hamarneh, 2016]. We focus in this chapter on regularization strategies for image segmentation where the objective is to enforce the connectivity or smoothness of the segmented structures in images. The investigation of spatial priors for image segmentation is performed within the framework of variational Bayesian methods and is motivated by the several challenges. The first challenge is to perform the inference of all model parameters in an efficient way that can scale easily with the image size. In particular, all proposed models lead to closed-form update formulas that can be easily implemented and contribute to the interpretability of those algorithms. The second challenge is to efficiently constrain the smoothness of recovered segmented structures in a data-driven way. In particular, it is expected that, depending on the image content, several levels of regularization can be achieved. The third challenge is to allow realistic uncertainty quantification both in terms of local and global segmentation measurements.

In this chapter we investigate six generic priors that allow spatial regularization. Five of these were proposed in previous works, while the sixth is based on a Gaussian process and, to the best of our knowledge, is novel. In particular, we compare the classical Markov random field (MRF) prior based on connectivity, and its conditional random field (CRF) variant, to its continuous counterpart, which penalizes the total variations (TV) of the prior label map or the squared norm of its derivatives. The latter will be denoted by FDSP (Finite Difference Spatial Prior) throughout the chapter. MRF priors are widespread and can be found for instance in [Held et al., 1997; Warfield et al., 2004; Xu et al., 2010]. A CRF prior is used by the well-known GrabCut algorithm [Rother et al., 2004] to regularize the segmentation, while [Bioucas-Dias & Figueiredo, 2016; Babacan et al., 2008; Babacan et al., 2009] are examples of works employing a TV prior. The FDSP prior was introduced in a previous work of the authors [Audelan & Delingette, 2020]. It is related to the Gauss-Markov random field prior developed in [Figueiredo, 2005b], which is a straightforward extension of the MRF prior to the continuous case and will not be analysed in this chapter.

Furthermore, we introduce a spatial prior defined through a Gaussian process, which has, to the best of our knowledge, never been proposed before. Optimization techniques are provided to address the well-known memory and computational burdens associated with the storage and inversion of the Gaussian process covariance matrix. In addition, we discuss its relation to another prior defined as a generalized linear model of basis functions spread over the image grid, denoted by GLSP (Generalized Linear Spatial Prior). The GLSP prior was also introduced in a previous work of the authors [Audelan

& Delingette, 2019]. A supplementary inference strategy is provided for the GLSP prior in this chapter, based on a Laplace approximation of the lower bound. We develop an approach which generalizes the fast relevance vector machine (RVM) proposed by Tipping and Faul [Tipping & Faul, 2003] and leads to an incremental algorithm capable of selecting the most relevant basis from a user defined dictionary. Thus, this extension allows the regularization to be automatically spatially adaptive in a data-driven fashion.

We restrict the analysis to binary image segmentation, although we show at the end of the chapter that the extension to multiple classes is straightforward. The appearance component of segmentation models is not the main focus of this study. In the remainder, we model the image likelihood of each region as mixtures of Gaussian distributions. The framework therefore corresponds to a mixture of mixtures model, known to increase the robustness with respect to outliers [Malsiner-Walli et al., 2017; Orbanz & Buhmann, 2005; Li et al., 2015]. Moreover, a Dirichlet prior is introduced over the mixing coefficients of the appearance models allowing the appropriate number of components to be automatically selected, leading to a more robust description.

A contribution of the chapter is to unify the aforementioned spatial priors into a common inference scheme based on variational calculus. Variational inference is based on the maximization of a lower bound over the data marginal likelihood and learns probability distributions, thus allowing uncertainty to be quantified. Convergence is also guaranteed, as each iteration corresponds to an increase in the data marginal likelihood. The lower bound is also a useful tool to perform model selection, by comparing the values reached after convergence. Finally, but also importantly, variational inference allows an automatic data-driven estimation of the parameters, including in some cases those controlling the level of regularization. In addition to performing classical mean-field approximations assuming factorized approximate posterior distributions, we also make use of local variational bounds [Bishop, 2006; Murphy, 2012] which ensure the tractability of closed-form updates despite using nonconjugated distributions. After convergence, the generative nature of the model allows new plausible segmentation samples to be generated.

Finally, we also introduce a formalization of the segmentation problem based on the evolution of a narrow band along the foreground boundaries at each iteration as in the level-set algorithm. The prior label map of the segmentation is initialized by a seed provided by the user. The algorithm alternates between two steps, one that estimates the appearance parameters and the other that performs the spatial regularization. The latter focuses at each iteration on the segmentation boundaries, leading to an evolution of the structure by shrinking or expansion. The process is repeated until convergence.

We summarize the main contributions of the chapter below:

- A common inference scheme based on variational inference is provided for several spatial priors.
- A novel spatial prior based on a Gaussian process is introduced, together with some optimization methods based on Fourier transform or Kronecker products allowing the approach to scale to the segmentation of large images.
- An incremental algorithm is proposed for the GLSP based on sparse Bayesian learning.
- A formulation is given of the algorithm, based on a narrow band evolution.

The rest of the chapter is organized as follows. Section 3.2 presents the probabilistic segmentation framework with the intensity model. In section 3.3, we present the different spatial priors and the optimization techniques required for the Gaussian process. Section 3.4 develops the common inference scheme, based on variational inference, and gives details specific to each spatial prior. Finally, the last section shows some results and comparisons using medical imaging data.

3.2 Probabilistic segmentation framework

We consider the segmentation of an image, I, made of N voxels with intensities $I_n \in \mathbb{R}^D$, $n = 1, \ldots, N$ into $K = 2$ regions. We introduce for each voxel a binary hidden random variable $Z_n \in \{0, 1\}$ with $Z_n = 1$ if voxel n belongs to the structure of interest.

Appearance models. The foreground and background regions are defined by the two image likelihoods $p(I_n|Z_n = 1, \theta_I^0)$ and $p(I_n|Z_n = 0, \theta_I^1)$, respectively, where θ_I^0 and θ_I^1 are parameters governing those models. At this point, two situations can arise [Figueiredo, 2005b]. In a supervised setting, the image likelihoods are given, for instance by neural networks, and the objective is to regularize the segmentation. On the other hand, the unsupervised setting aims to learn the intensity parameters θ_I jointly with the level of regularization. In fact, the supervised setting can be considered as a special case of the second one, with the intensity parameters being fixed. In this chapter, we will consider parametric appearance models based on mixtures of Gaussian distributions:

$$p(I_n|Z_{nk} = 1) = \sum_{m=1}^{M_k} \pi_{km} \mathcal{N}(I_n; \mu_{km}, \mathbf{\Lambda}_{km}^{-1}) \,, \tag{3.1}$$

where M_k is the number of mixture components in region k and π_k are the mixing coefficients, with the constraint $\sum_{m=1}^{M_k} \pi_{km} = 1$. We introduce a new categorical variable Γ which is a binary 1-of-M_k encoding such that $\Gamma_{nkm} = 1$ if voxel n belongs to the m-th component of region k, and $\sum_m \Gamma_{nkm} = 1$. It leads to the following likelihood:

$$p(I_n|\Gamma_n, Z_n) = \prod_{k=1}^{K} \left[\prod_{m=1}^{M_k} \mathcal{N}(I_n; \mu_{km}, \mathbf{\Lambda}_{km}^{-1})^{\Gamma_{nkm}} \right]^{Z_{nk}} . \tag{3.2}$$

In order to obtain a robust description, a sparsity-inducing Dirichlet prior is chosen over the mixing coefficients π:

$$p(\pi) = \prod_{k=1}^{K} \mathrm{Dir}(\pi_k; \gamma_{k0}) = \prod_{k=1}^{K} \left\{ C(\gamma_{k0}) \prod_{m=1}^{M_k} \pi_{km}^{\gamma_{k0}-1} \right\} , \tag{3.3}$$

where $C(\gamma)$ is the normalization constant. Depending on the value of the hyperparameter γ_0, this prior will favor sparse representation allowing unnecessary components to be removed from the model. In order to get a fully Bayesian approach, the mean and

precision of each mixture component are equipped with a Gaussian-Wishart prior such that:

$$\begin{aligned} p(\mu, \mathbf{\Lambda}) &= p(\mu|\mathbf{\Lambda})p(\mathbf{\Lambda})\,, \\ &= \prod_{k=1}^{K}\prod_{m=1}^{M_k} \mathcal{N}(\mu_{km}; m_{k0}, (\beta_{k0}\mathbf{\Lambda}_{km})^{-1})\mathcal{W}(\mathbf{\Lambda}_{km}; W_{k0}, \nu_{k0})\,. \end{aligned} \tag{3.4}$$

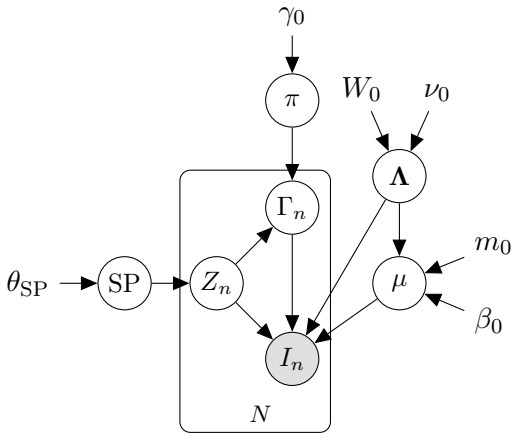


Fig. 3.1: Graphical model of the probabilistic framework. SP denotes the spatial prior and θ_{SP} its hyperparameters.

Fig. 3.1 shows a graphical representation of the probabilistic framework. SP denotes the spatial prior introduced to impose smoothness over the label map Z. The different spatial priors and their hyperparameters are presented in the next section.

3.3 Spatial smoothness priors

Spatial priors are required to enforce the spatial consistency of the segmented structure. In this chapter, four classes of spatial priors are investigated.

3.3.1 MRF/CRF priors

A classical approach to enforce spatial smoothness is to define a Markov random field (MRF) prior over the label map. Its simple and straightforward formulation made it widely popular, for example [Held et al., 1997; Warfield et al., 2004; Woolrich et al., 2009; Xu et al., 2010]. This prior is essentially based on a discrete representation of an image where pixels are connected through a graph. It thus enforces the connectivity among voxels and reflects the prior assumption that it is likely that two neighboring voxels also share the same label. A natural way to represent the interactions between neighboring voxels is the Ising model for the binary case, extended to the multi-class setting by the Potts model. The label prior probability is then defined as:

$$p(Z|\beta) = \frac{1}{T(\beta)} \exp\left(\frac{\beta}{2}\sum_{n=1}^{N}\sum_{i\in\delta(n)}\sum_{k=0}^{1} Z_{nk}Z_{ik}\right)\,, \tag{3.5}$$

where $\beta \geq 0$ is the hyperparameter controlling the strength of the regularization and $T(\beta)$ is the partition function. $\delta(n)$ are the neighboring voxels of n. A β value close to zero amounts to a model without regularization, while large values encourage neighboring voxels to belong to the same region.

The main limitation of the MRF is the intractability of the normalization constant $T(\beta)$, needed for the automatic estimation of the hyperparameter, as it requires considering all possible configurations of the MRF, which is computationally impossible for large lattices. Adaptive estimation of the hyperparameter β is an open field of research [Woolrich et al., 2005; Woolrich & Behrens, 2006; Pereyra & McLaughlin, 2017], but in this chapter we restrict the analysis to the case where β is given by the user.

The MRF prior relies solely on connectivity assumptions and does not take image features into account. Conditional random fields (CRF) are a variant of MRF priors that incorporate intensity information. Thus the prior probability no longer depends only on the label of neighboring voxels, but also on the variation in intensity across these voxels. The contrast-sensitive prior can be written as [Boykov & Jolly, 2001; Rother et al., 2004]:

$$p(Z|\beta) = \frac{1}{T(\beta,\gamma)} \exp\left(\frac{\beta}{2}\sum_{n=1}^{N}\sum_{i\in\delta(n)}\sum_{k=0}^{1} Z_{nk}Z_{ik}\frac{\exp\left(-\gamma(I_n - I_i)^2\right)}{\text{dist}(n,i)}\right), \tag{3.6}$$

where $dist(x, y)$ is the Euclidean distance between voxels. $\gamma \geq 0$ is a new hyperparameter controlling the importance of the intensity part of the prior: with $\gamma = 0$ the prior is identical to Eq. 3.5. Again, γ cannot be learnt automatically and is fixed to $\gamma = (2\langle(I_i - I_j)^2\rangle)^{-1}$, where $\langle.\rangle$ is the expectation over the image [Rother et al., 2004].

3.3.2 TV/FDSP priors

The MRF and CRF priors are defined directly on the label field Z. The discrete nature of the variable makes it more difficult to manipulate and leads to complex combinatorial optimization problems, in particular for the computation of the partition function. Another approach is to introduce a new continuous hidden variable $\mathbf{W} = [w_1, \ldots, w_N]^T$ related to the label via the sigmoid function $\sigma(x) = 1/(1 + e^{-x})$:

$$p(Z|\mathbf{W}) = \prod_{n=1}^{N} [\sigma(w_n)]^{Z_n} [\sigma(-w_n)]^{1-Z_n}. \tag{3.7}$$

Priors defined over $\mathbf{W}$ can then be used to enforce spatial consistency. In [Bioucas-Dias & Figueiredo, 2016; Babacan et al., 2008; Babacan et al., 2009], the prior penalizes the total variation (TV) of the continuous field:

$$p(\mathbf{W}|\alpha) = \frac{1}{T(\alpha)} \exp\left(-\alpha\sum_{n=1}^{N}\sqrt{\sum_{d=1}^{D}(w_n - w_{\delta_d(n)})^2}\right), \tag{3.8}$$

where α controls the amount of spatial regularization and $\delta_d(n)$ denotes the nearest neighbor of voxel n in the dimension d. For a two-dimensional image ($D = 2$), $\delta_d(n)$ represents the first neighbor to the left or above for $d = 0$ and $d = 1$, respectively. The function $h(x) = \sum_n \sqrt{\sum_d (x_n - x_{\delta_d(n)})^2}$ is 1-homogenous and therefore the normalization

factor can be written $T(\alpha) = c\alpha^{-N}$, where c does not depend on α [Pereyra et al., 2015]. This prior is known for not over-penalizing discontinuities in the image while enforcing spatial coherence [Babacan et al., 2009].

However, a downside of the TV approach is the presence of the square root function requiring special treatment to ensure tractability. In [Audelan & Delingette, 2020], we proposed another prior imposing smoothness by penalizing the squared norm of the derivatives of order p of the vector $\mathbf{W}$. It is denoted by Finite Difference Spatial Prior (FDSP) and is written:

$$p(\mathbf{W}|\alpha) = \frac{1}{T(\alpha)} \exp\left(-\alpha \sum_n ||\Delta_p(w_n)||^2\right) . \tag{3.9}$$

$\Delta_p(w_n)$ is the central finite difference operator of order p at w_n. In this paper, we consider only first order derivatives ($p = 1$) and the prior is written:

$$p(\mathbf{W}|\alpha) = \frac{1}{T(\alpha)} \exp\left(-\frac{\alpha}{4} \sum_{n=1}^{N} \sum_{d=1}^{D} (w_{\delta_d(n+1)} - w_{\delta_d(n-1)})^2\right) , \tag{3.10}$$

where $\delta_d(n+i)$ represents the neighbor with index i of voxel n in the dimension d. The quantity $h(x) = \sum_n ||\Delta_p(w_n)||^2$ is 2-homogenous and the normalization factor has the form $T(\alpha) = c\alpha^{-N/2}$ [Pereyra et al., 2015]. Another way to see this is to note that $p(\mathbf{W}|\alpha)$ is a zero mean Gaussian distribution whose precision matrix $\mathbf{\Lambda}_{\text{FDSP}}$ consists of difference operators. Therefore, in contrast to the TV prior, the normalization constant can be fully expressed, leading to $T(\alpha) = (4\pi)^{N/2}\alpha^{-N/2}|\mathbf{\Lambda}_{\text{FDSP}}|^{-1/2}$ for the case $p = 1$.

3.3.3 GP prior

As in the previous section, we use a continuous hidden field $\mathbf{W}$ to enforce the spatial consistency of the label variable, leading to the same label prior $p(Z|\mathbf{W}) = \prod_{n=1}^{N} [\sigma(w_n)]^{Z_n} [\sigma(-w_n)]^{1-Z_n}$. However, we replace the TV/FDSP priors penalizing the continuous field by a regularizer formulated as a Gaussian process. Gaussian processes (GP) are a generalization of Gaussian multivariate distributions [Rasmussen & Williams, 2005]. They provide a framework to define a prior over spatially correlated variables in a straightforward manner. A zero mean Gaussian prior is chosen for the vector $\mathbf{W}$ with a covariance matrix encoding the spatial relationships between voxels:

$$p(\mathbf{W}|\theta_{\text{GP}}) = \mathcal{N}(\mathbf{W}; 0, \mathbf{\Sigma}_{\text{GP}}) . \tag{3.11}$$

In this chapter, we consider the squared exponential function with hyperparameters ω_0 and ω_1:

$$\forall\, a, b \in \mathbb{R}^D,\ \mathbf{\Sigma}_{\text{GP}}(a, b) = \omega_0^2 \exp\left(-\frac{||a-b||^2}{\omega_1^2}\right) . \tag{3.12}$$

ω_0 controls the amount of variability in the continuous field defined by $\mathbf{W}$, while ω_1 is the characteristic length scale representing the typical correlation length between two voxels.

A well-known difficulty of GP is their poor scalability to high dimensional datasets, in particular images. The size of the covariance matrix (N^2) makes the GP computationally

and memory intensive at the same time [Rasmussen & Williams, 2005]. However, optimization techniques exist when the covariance matrix is stationary (i.e., invariant by translation) and when the points lie on a regular grid, which is the case for images.

A first approach assumes periodic boundary conditions on the image [Kozintsev, 1999]. $\mathbf{\Sigma}_{\text{GP}}$ is then a symmetric Block Circulant matrix with Circulant Blocks (BCCB matrix) such that each column is a periodic shift of the first column. The first column contains all the information which limits the storage to a vector of size N. Moreover, theoretical results on BCCB matrices allow us to write $\mathbf{\Sigma}_{\text{GP}} = F^{-1}\Delta F$, where Δ is the diagonal matrix of eigenvalues and F is the $N \times N$ discrete Fourier transform matrix [Kozintsev, 1999]. Computations involving $\mathbf{\Sigma}_{\text{GP}}$ are then performed very efficiently in the Fourier domain.

The periodic boundary conditions assumption can be a limitation, especially for small images or strong spatial regularization. In these cases, another approach based on the separability of the covariance function may be more suitable [Saatchi, 2011]. The squared exponential kernel can be factorized along the image dimensions:

$$\forall\, a, b \in \mathbb{R}^D,\ \mathbf{\Sigma}_{\text{GP}}(a, b) = w_0 \prod_{d=1}^{D} \exp\left(-\frac{||a_d - b_d||^2}{\omega_1^2}\right) . \tag{3.13}$$

The GP matrix can then be written as a Kronecker product of the D covariance matrices along each dimension: $\mathbf{\Sigma}_{\text{GP}} = \bigotimes_{d=1}^{D} \mathbf{\Sigma}_{\text{GP}}^d$. In practice, the full GP matrix is never constructed and only the smaller $\mathbf{\Sigma}_{\text{GP}}^d$ of sizes $N_d \times N_d$ are stored. Moreover, an efficient framework was developed in [Saatchi, 2011] for fast computations of matrix/vector products. This optimization approach is valid for any length-scale values, including large ones, thus allowing long range correlations to be efficiently taken into account. It is an advantage compared to the MRF/CRF priors, for which the modelling of large-range interactions involves the consideration of larger neighborhoods and thus leads to increased complexity [Bouman & Shapiro, 1994].

3.3.4 GLSP prior

The last prior investigated in this chapter was introduced in a previous work of the authors [Audelan & Delingette, 2019] and is denoted by Generalized Linear Spatial Prior (GLSP). The label prior is still defined as a Bernoulli distribution, but its parameter is now a spatially random function specified as a generalized linear model:

$$p(Z_n = 1|\mathbf{W}) = \sigma\left(\sum_{l=1}^{L} \Phi_l(\mathbf{x}_n) w_l\right) , \tag{3.14}$$

where $\mathbf{x}_n \in \mathbb{R}^D$ denotes the position of voxel n. The basis $\{\Phi_l(\mathbf{x})\}$ are L functions of space, typically radial basis functions (for instance, Gaussian functions) defined on a regular grid. Each basis function has an associated weight $w_l \in \mathbf{W}$, now of size L. It is clear that the prior probabilities of two geometrically close voxels are related to each other through the smoothness of the label field parameterized by the function $f(\mathbf{x}_n) = \sum_{l=1}^{L} \Phi_l(\mathbf{x}_n) w_l = \mathbf{\Phi}_n^T \mathbf{W}$, writing $\mathbf{\Phi}_n^T = [\Phi_1(x_n), \cdots, \Phi_L(x_n)]$. The L basis functions $\{\Phi_l(\mathbf{x})\}$ are commonly uniformly spread over the image domain and their choice influences the strength of the spatial regularization. The key parameters are the spacing s between the basis centers, the standard deviations (or radii) r of the Gaussian

functions and the position of the origin basis. Together, they influence the amount of smoothing produced by the label prior, large spacing and standard deviations leading to smoother prior probability maps.

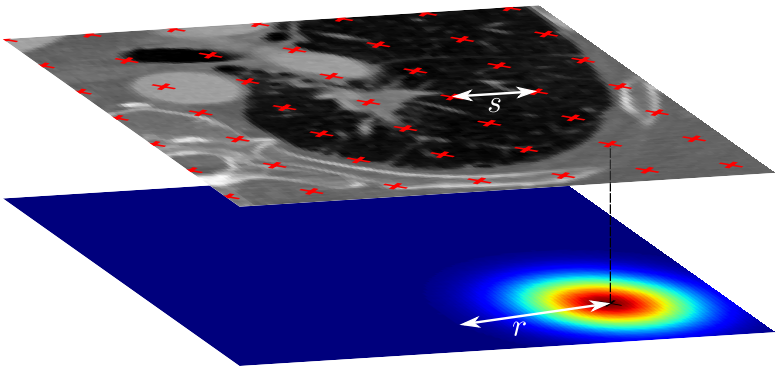


Fig. 3.2: Example of a grid of basis functions used as input by the GLSP prior.

The level of regularization is also controlled by the prior defined over the weights. To obtain a robust description, the vector $\mathbf{W}$ is equipped with a prior of the form:

$$p(\mathbf{W}|\alpha) \propto \exp\left(-\frac{\alpha}{2}\mathbf{W}^T\mathbf{R}\mathbf{W}\right) . \tag{3.15}$$

Depending on the precision matrix $\mathbf{R}$, a wide variety of priors can be encoded. For instance, by setting $\mathbf{R} = \mathbf{I}_L$, we obtain a zero mean Gaussian prior with precision matrix $\alpha\mathbf{I}_L$. It constrains the values of $\mathbf{W}$ by penalizing the vector norm. Alternatively, $\mathbf{R}$ can be designed such that it penalizes the magnitude $||f||^2$ of the prior label field or of its derivatives [Le Folgoc et al., 2017] using reproducing kernel Hilbert space (RKHS) methods. Details of the computation of the coefficients of the matrix $\mathbf{R}$ can be found in appendix A.3.

In addition, a non-informative uniform prior is chosen for the hyperparameter α, i.e., $p(\alpha) \propto 1$.

Finally, note that it is possible to establish a connection between the GLSP prior and the one based on a GP. If $\mathbf{R}$ is chosen as the identity matrix, the GLSP prior is then just a transfer of the Relevance Vector Machine (RVM) framework [Tipping & Faul, 2003] for binary classification to the image segmentation setting. In the case of a noise-free regression model, one can write the RVM as a GP after marginalizing out the weight variable [Tipping, 2001; Rasmussen & Williams, 2005]. The covariance function is then $k(a,b) = \sum_{l=1}^{L} \alpha^{-1}\phi_l(a)\phi_l(b)$ and clearly has a finite number of nonzero eigenvalues, meaning that it is degenerate. On the contrary, the squared exponential kernel is nondegenerate [Rasmussen & Williams, 2005] with an associated feature space of infinite dimension. It confers higher expressiveness to the GP model than is the case in the basis functions framework, which can suffer from limited flexibility. However, the latter is more efficient in terms of computation and memory, because usually $L << N$.

The graphical model associated with each spatial prior is given in Fig. 3.3.

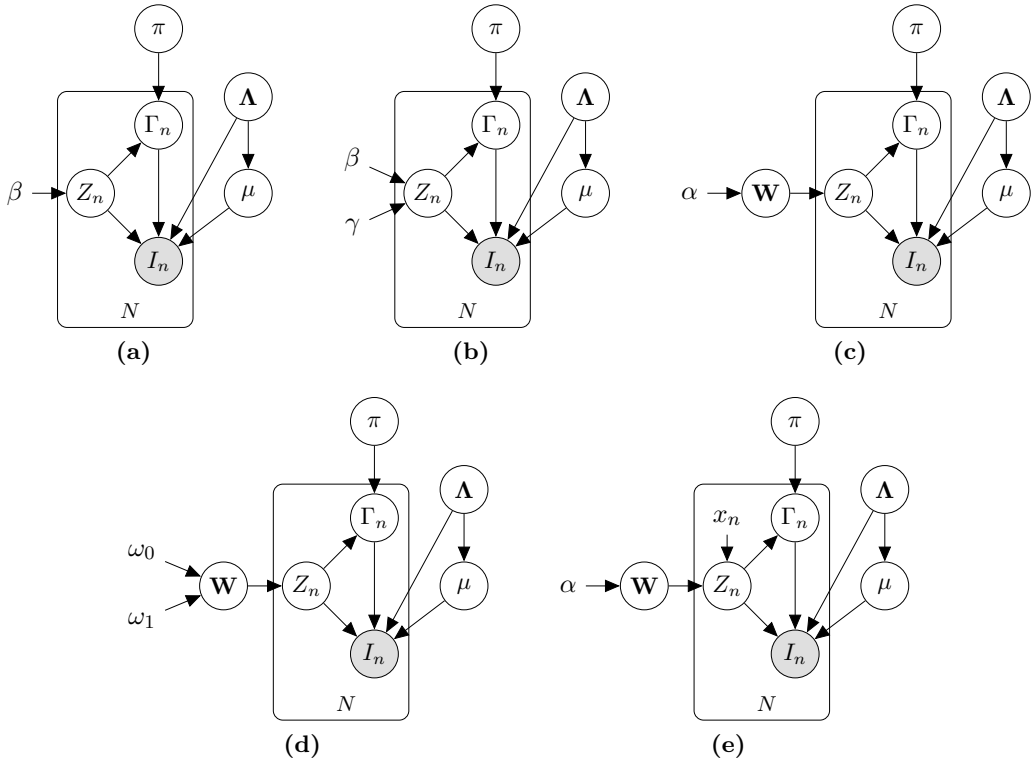


Fig. 3.3: Graphical models of the probabilistic framework with a MRF prior (3.3a), a CRF prior (3.3b) a TV or FDSP priors (3.3c), a GLSP prior (3.3e) and finally a GP prior (3.3d). The prior parameters of the appearance variables are not shown to keep the representation uncluttered.

3.4 Model inference

We propose a common inference framework based on variational calculus allowing the appearance and regularization parameters $U = \{Z, \Gamma, \mathbf{W}, \mu, \mathbf{\Lambda}, \pi\}$ to be learnt. (For an MRF or CRF prior, the variable $\mathbf{W}$ is ignored.) The objective is to maximize the data log marginal likelihood, $\log p(I)$, while approximating the true posterior distribution, $p(U|I)$, by a chosen family of distributions, $q(U)$. An increase in the data log likelihood can be achieved by minimizing the Kullback-Leibler divergence between the true posterior, $p(U|I)$, and the approximation, $q(U)$, or equivalently by maximizing the lower bound, $\mathcal{L}(q)$:

$$\log p(I) = \underbrace{\int_U q(U) \log \frac{p(I,U)}{q(U)}}_{\mathcal{L}(q)} + \underbrace{\mathrm{KL}[q(U)||p(U|I)]}_{\geq 0} . \tag{3.16}$$

We further assume that the approximation of the posterior factorizes with respect to each variable (mean field approximation) such that:

$$q(U) = q(Z)q(\Gamma)q(\mathbf{W})q(\mu, \mathbf{\Lambda})q(\pi) . \tag{3.17}$$

With $\theta_I = \{\mu, \mathbf{\Lambda}, \pi\}$ gathering the intensity variables, the lower bound can then be rewritten as:

$$\begin{aligned}\log p(I) \geq \mathcal{L}(q) = &\sum_{Z} \sum_{\Gamma} \int_{\mathbf{W}} \int_{\theta_I} q(Z)q(\Gamma)q(\mathbf{W})q(\theta_I) \\ &\log \frac{p(I, Z, \Gamma, \mathbf{W}, \theta_I)}{q(Z)q(\Gamma)q(\mathbf{W})q(\theta_I)} d\mathbf{W} d\theta_I \,.\end{aligned} \tag{3.18}$$

If q_i denotes any of the factors in Eq. 3.17 and q_{-i} the product of the remaining factors, we know by variational calculus that the distribution q_i^* maximizing Eq. 3.18 has the form:

$$\log q_i^* = \mathbb{E}_{q_{-i}}[\log p(I, U)] + cst \,, \tag{3.19}$$

when fixing the other distributions q_{-i}. This results leads to an iterative algorithm where the lower bound is optimized with respect to each approximate distribution q_i in turn. In the following sections we present the main results for each of the posterior distribution approximations; details of the derivations can be found in appendix A.1.

3.4.1 Label posterior approximation

The update of the label posterior approximation is equivalent to the expectation step in an Expectation-Maximization (EM) algorithm. Eq. 3.19 applied to $q(Z_n)$ leads to a Bernoulli distribution of parameters η_{n0} and η_{n1} with $\eta_{nk} = \rho_{nk} / \sum_k \rho_{nk}$ for $k \in \{0, 1\}$ and:

$$\begin{aligned}\log \rho_{nk} = \sum_{m=1}^{M_k} \mathbb{E}[\Gamma_{nkm}] \Big[&-\frac{D}{2} \log(2\pi) + \frac{1}{2} \mathbb{E}[\log |\mathbf{\Lambda}_{km}|] \\ &-\frac{1}{2} \mathbb{E}[(I_n - \mu_{km})^T \mathbf{\Lambda}_{km} (I_n - \mu_{km})] + \mathbb{E}[\log \pi_{km}] \Big] + R_{nk} \,.\end{aligned} \tag{3.20}$$

R_{nk} is the regularization term that varies depending on the spatial prior:

- $R_{nk} = \beta \sum_{j \in \delta(n)} \eta_{jk}$ for the MRF prior.
- $R_{nk} = \beta \sum_{j \in \delta(n)} \eta_{jk} \exp\left(-\gamma (I_n - I_j)^2\right) / \text{dist}(n, j)$ for the CRF prior.
- $R_{nk} = k \mathbb{E}[w_n]$ for the TV, FDSP and GP priors.
- And finally, $R_{nk} = k \mathbb{E}[\mathbf{\Phi}_n^T \mathbf{W}]$ for the GLSP prior.

The update for the MRF and CRF priors involves a fixed-point equation, the new values η_n^{i+1} at iteration $i+1$ are obtained from the values of the same Bernoulli distribution parameters at iteration i.

3.4.2 Appearance parameters

This section gives the posterior approximations for the variables Γ, π, μ and $\mathbf{\Lambda}$. We chose to model the appearance of both regions using mixtures of Gaussian distributions and Γ represents the component of each voxel in each region. The posterior approximation

$q(\Gamma_{nk})$ is a multinomial distribution with parameters $\delta_{nkm} = \tau_{nkm} / \sum_m \tau_{nkm}$ for $1 \leq m \leq M_k$ and:

$$\begin{aligned} \log \tau_{nkm} = \eta_{nk} \Big[& -\frac{D}{2} \log(2\pi) + \frac{1}{2} \mathbb{E}[\log |\mathbf{\Lambda}_{km}|] \\ & -\frac{1}{2} \mathbb{E}[(I_n - \mu_{km})^T \mathbf{\Lambda}_{km} (I_n - \mu_{km})] + \mathbb{E}[\log \pi_{km}] \Big] . \end{aligned} \tag{3.21}$$

Updates for the remaining variables are classical results for variational mixtures of Gaussian distributions [Bishop, 2006]. $q(\pi_k)$ is thus a Dirichlet distribution $\mathrm{Dir}(\pi_k; \gamma_k)$, $\gamma_k = \{\gamma_{k1}, \ldots, \gamma_{kM_k}\}$ and $\gamma_{km} = \sum_{n=1}^{N} \eta_{nk} \delta_{nkm} + \gamma_{k0} = N_{km} + \gamma_{k0}$. $q(\mu_{km}, \mathbf{\Lambda}_{km})$ follows a Gaussian-Wishart distribution $\mathcal{N}(\mu_{km}; m_{km}, (\beta_{km} \mathbf{\Lambda}_{km})^{-1}) \mathcal{W}(\mathbf{\Lambda}_{km}; W_{km}, \nu_{km})$, with parameters given by:

$$\beta_{km} = \beta_0 + N_{km} , \tag{3.22}$$

$$m_{km} = \frac{1}{\beta_{km}} (\beta_0 m_0 + N_{km} \overline{I}_{km}) , \tag{3.23}$$

$$W_{km}^{-1} = W_0^{-1} + N_{km} S_{km} + \frac{\beta_0 N_{km}}{\beta_0 + N_{km}} (\overline{I}_{km} - m_0)(\overline{I}_{km} - m_0)^T , \tag{3.24}$$

$$\nu_{km} = \nu_0 + N_{km} . \tag{3.25}$$

$\overline{I}_{km}$ and S_{km} are defined as:

$$\overline{I}_{km} = \frac{1}{N_{km}} \sum_{n=1}^{N} \eta_{nk} \delta_{nkm} I_n , \tag{3.26}$$

$$S_{km} = \frac{1}{N_{km}} \sum_{n=1}^{N} \eta_{nk} \delta_{nkm} (I_n - \overline{I}_{km})(I_n - \overline{I}_{km})^T . \tag{3.27}$$

3.4.3 Regularization variables

In this section, we propose some strategies to learn the posterior approximation $q(\mathbf{W})$ of the spatial regularization variable. It does not concern the MRF and CRF priors, as in these cases the regularization is applied directly on the discrete variable Z and not through the continuous variable $\mathbf{W}$.

If we apply Eq. 3.19 to $\mathbf{W}$, we get $\log q^*(\mathbf{W}) = \mathbb{E}\left[\log \left(P(Z|\mathbf{W}) p(\mathbf{W})\right)\right]$. A first problem common to all priors arises with the expectation taken over the label prior. Indeed, $p(Z|\mathbf{W})$ is a Bernoulli distribution whose parameters involve the sigmoid function, making the expectation intractable. Regarding the expectation over the continuous variable $\mathbf{W}$, it involves a Gaussian distribution for the FDSP, GP and GLSP priors and is easy to compute. However, there is an issue with the square root function in the TV prior which makes the integral intractable.

To obtain a tractable approximation, we propose to use an approach based on local variational bounds as an alternative to the Laplace approximation. A new bound is introduced over the problematic distributions. The objective of maximizing the data log likelihood then becomes one of optimizing a new lower bound over the lower bound $\mathcal{L}(q)$.

Lower bound over the square root function

We follow the approach of [Babacan et al., 2008; Babacan et al., 2009] who use the following bound over the square root function:

$$\forall x \geq 0, \forall y > 0, \quad \sqrt{x} \leq \frac{x+y}{2\sqrt{y}} \,. \tag{3.28}$$

The initial formulation of the TV prior is then replaced by $L(\mathbf{W}, \alpha, u)$ according to the following inequality:

$$p(\mathbf{W}|\alpha) \geq L(\mathbf{W}, \alpha, u) = c\alpha^N \exp\left(-\frac{\alpha}{2}\sum_{n=1}^{N}\frac{\sum_{d=1}^{D}(w_n - w_{\delta_d(n)})^2 + u_n}{\sqrt{u_n}}\right) \,. \tag{3.29}$$

The prior is now a Gaussian distribution and the expectation $\mathbb{E}[\log p(\mathbf{W}|\alpha)]$ is now tractable. The variables u are additional variational parameters that can be estimated.

JJ bound over the label prior

The intractability problem is caused by the sigmoid function in the parameters of the Bernoulli distribution. We follow the approach introduced by [Jaakkola & Jordan, 2000] in the context of logistic regression and replace the sigmoid function according to this inequality: $\sigma(x) \geqslant l(x,\xi) = \sigma(\xi)\exp\left[(x-\xi)/2 - \lambda(\xi)(x^2-\xi^2)\right]$. ξ is an additional variational parameter and $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$. The spatial prior $p(Z|\mathbf{W})$ can thus be approximated by $F(Z, \mathbf{W}, \xi) = \prod_n [l(y_n, \xi_n)]^{Z_n} [l(-y_n, \xi_n)]^{1-Z_n}$, with $y_n = \mathbf{\Phi}_n^T \mathbf{W}$ for the GLSP prior and w_n otherwise. This approach will be referred to as the JJ bound after its inventors, as done in [Murphy, 2012].

This approximation leads to a new lower bound $\mathcal{J}(q)$ on the lower bound $\mathcal{L}(q)$:

$$\begin{aligned}\log p(I) \geqslant \mathcal{L}(q) \geqslant \mathcal{J}(q) = \sum_Z \sum_\Gamma \int_{\mathbf{W}} \int_{\theta_I} q(Z)q(\Gamma)q(\mathbf{W})q(\theta_I) \\ \log \frac{p(I|\Gamma, Z, \theta_I)p(\Gamma|Z, \theta_I)p(\theta_I)F(Z, \mathbf{W}, \xi)p(\mathbf{W})}{q(Z)q(\Gamma)q(\mathbf{W})q(\theta_I)} d\mathbf{W} d\theta_I \,.\end{aligned} \tag{3.30}$$

$p(\mathbf{W})$ is further replaced by $L(\mathbf{W}, \alpha, u)$ for the TV prior. The new lower bound $\mathcal{J}(q)$ is now tractable. Note that the use of the JJ bound is possible here because we consider in this chapter binary segmentation problems and the sigmoid function is a classical way to define the parameters of the Bernoulli distribution. In the multi-class setting however, we would use a multinomial logistic regression framework with the softmax function $\sigma(x)_i = e^{x_i}/\sum_j e^{x_j}$ which also causes tractability issues. Nonetheless, we can keep the local variational bound approach using a bound on the LogSumExp function, for example, the Böhning bound [Böhning, 1992].

Applied to the GLSP prior, the JJ bound leads to a Gaussian distribution $\mathcal{N}(\mathbf{W}; \mu_{\mathbf{W}}, \mathbf{\Sigma}_{\mathbf{W}})$ for $q(\mathbf{W})$ with parameters given by:

$$\mathbf{\Sigma}_{\mathbf{W}} = \left[\mathbf{\Phi}\mathbf{B}\mathbf{\Phi}^T + \alpha\mathbf{R}\right]^{-1} \,, \tag{3.31}$$

$$\mu_{\mathbf{W}} = \mathbf{\Sigma}_{\mathbf{W}} \sum_{n=1}^{N} \left(\eta_{n1} - \frac{1}{2}\right) \mathbf{\Phi}_n \,. \tag{3.32}$$

where $\mathbf{\Phi}$ is the $L \times N$ matrix obtained by stacking the vectors $\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_N$ and $\mathbf{B} = \text{diag}([2\xi_1, \ldots, 2\xi_N])$. For the TV and FDSP priors, we further assume the factorization $q(\mathbf{W}) = \prod_n q(w_n)$. The variational optimization for $q(w_n)$ yields a normal distribution $\mathcal{N}(w_n; \mu_{w_n}, \mathbf{\Sigma}_{w_n})$. The parameters for the TV prior are given by:

$$\mathbf{\Sigma}_{w_n} = \left[2\lambda(\xi_n) + \sum_d \frac{\alpha}{\sqrt{u_n}} + \sum_d \frac{\alpha}{\sqrt{u_{\xi_{d(n)}}}} \right]^{-1} , \tag{3.33}$$

$$\mu_{w_n} = \mathbf{\Sigma}_{w_n} \left[\eta_{n1} - \frac{1}{2} + \frac{\alpha}{\sqrt{u_n}} \sum_d \mathbb{E}[w_{\delta_{d(n)}}] + \alpha \sum_d \frac{\mathbb{E}[w_{\xi_{d(n)}}]}{\sqrt{u_{\xi_{d(n)}}}} \right] . \tag{3.34}$$

$\xi_{d(n)}$ denotes the voxel whose neighbor in dimension d is the voxel n. Likewise, we get a Gaussian distribution for the FDSP prior with parameters as follows:

$$\mathbf{\Sigma}_{w_n} = \left[2\lambda(\xi_n) + 2 \sum_d \frac{\alpha}{2} \right]^{-1} , \tag{3.35}$$

$$\mu_{w_n} = \mathbf{\Sigma}_{w_n} \left[\eta_{n1} - \frac{1}{2} + \frac{\alpha}{2} \sum_d \left(\mathbb{E}[w_{\delta_d(n+2)}] + \mathbb{E}[w_{\delta_d(n-2)}] \right) \right] . \tag{3.36}$$

We recall that for the FDSP prior $\delta_d(n+i)$ represents the neighbor of index i of voxel n in the dimension d.

Finally, we get also a normal distribution for the GP prior. The covariance $\mathbf{\Sigma_W}$ is written $\left[\mathbf{B} + \mathbf{\Sigma}_{\text{GP}}^{-1} \right]^{-1}$, where $\mathbf{B} = \text{diag}([2\xi_1, \ldots, 2\xi_N])$ is defined as above. Its size is $N \times N$, which is much larger than the $L \times L$ GLSP covariance matrix. To be able to handle such large matrices, we proposed in section 3.3.3 two optimization techniques based on Kronecker products or periodic boundary conditions. Using one of these methods, $\mathbf{\Sigma}_{\text{GP}}^{-1}$ can be computed and stored efficiently. However, the computation of $\mathbf{\Sigma_W}$ adds a nonconstant perturbation to the main diagonal, which leads to the loss of the BCCB structure or prevents a factorization in Kronecker products. We therefore propose to use another bound for the GP prior, in order to be able to use the optimization methods.

Böhning bound for the GP prior

The Böhning bound [Böhning, 1992] is a bound over the LogSumExp function $\text{lse}(x) = \log(1 + e^x)$ as follows: $\text{lse}(x) \leq ax^2/2 - bx + c$, where $a = 1/4$, $b = a\xi - g(\xi)$ and $c = a\xi^2/2 - g(\xi)\xi + \text{lse}(\xi)$. ξ is again an additional variational parameter and g is the gradient of the LogSumExp function. The Böhning bound is plotted with the JJ bound in Fig. 3.4. The JJ bound is tighter than the Böhning bound [Murphy, 2012], but here we are interested in the fact that the coefficient associated with the quadratic term is a constant, whereas it depended on ξ for the JJ bound, which will allow the optimization methods to be used. $q(\mathbf{W})$ is then also a Gaussian distribution, but the parameters are this time:

$$\mathbf{\Sigma_W} = \left[\frac{1}{4} \mathbf{I}_N + \mathbf{\Sigma}_{\text{GP}}^{-1} \right]^{-1} , \tag{3.37}$$

$$\mu_{\mathbf{W}} = \mathbf{\Sigma_W} \left[\eta_1 + B \right] , \tag{3.38}$$

where $\eta_1 = [\eta_{n1}, \ldots, \eta_{N1}]^T$ and $B = [b_1, \ldots, b_N]^T$. Taking the inverse of a matrix preserves the BCCB structure, as well as adding a constant value to the diagonal [Kozintsev, 1999]. If periodic boundary conditions are assumed, $\mathbf{\Sigma_W}$ is then also a BCCB matrix. Only the eigenvalues are stored and computations are made in the Fourier domain. If we use the second optimization method, we rely on the fact that $\mathbf{\Sigma}_{\text{GP}} = (\bigotimes_d Q_d)\, \Delta \left(\bigotimes_d Q_d^T\right)$, where Δ is the diagonal matrix of eigenvalues. As the perturbation of the diagonal is constant, it is possible to write $\mathbf{\Sigma_W} = (\bigotimes_d Q_d)\, [1/4\mathbf{I}_N + \Delta^{-1}]^{-1} \left(\bigotimes_d Q_d^T\right)$. Again, only the eigenvalues are stored and matrix/vector products involving $\mathbf{\Sigma_W}$, for instance Eq. 3.38, are computed efficiently following [Saatchi, 2011].

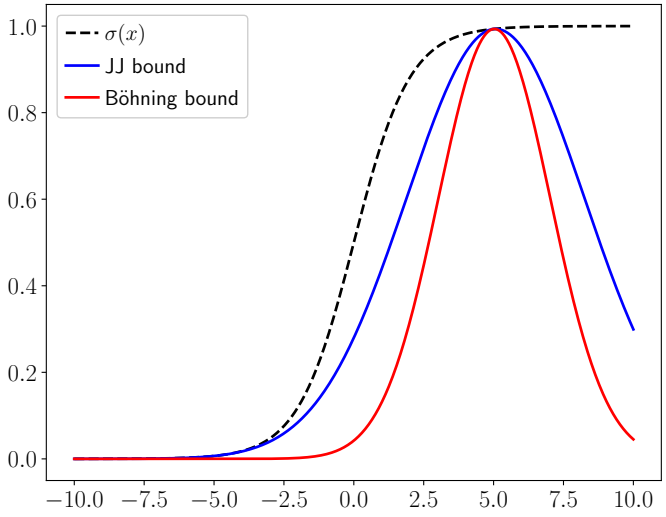


Fig. 3.4: Visualization of the JJ and Böhning bounds with respect to the sigmoid function, for $\xi = 5$. The JJ bound is tighter than the Böhning bound.

Update of prior parameters

The prior parameters are α for the TV, FDSP and GLSP priors and ω_0 and ω_1 for the GP prior. Assuming $q(\alpha)$ to be a Dirac distribution, we apply Eq. 3.19 to $q(\alpha)$ and take the derivatives which leads to the following update formula:

$$\alpha = \frac{L}{\mathbb{E}[\mathbf{W}^T \mathbf{R} \mathbf{W}]} \text{ for the GLSP prior,} \tag{3.39}$$

$$\alpha^{-1} = \frac{1}{2N} \sum_{n=1}^{N} \frac{\sum_{d=1}^{D} \mathbb{E}[(w_n - w_{\delta_{d(n)}})^2] + u_n}{\sqrt{u_n}} \text{ for the TV prior,} \tag{3.40}$$

$$\alpha^{-1} = \frac{1}{2N} \sum_{n=1}^{N} \sum_{d=1}^{D} \mathbb{E}\left[(w_{\delta_d(n+1)} - w_{\delta_d(n-1)})^2\right] \text{ for the FDSP prior.} \tag{3.41}$$

We could use the same approach to optimize the parameters of the GP covariance matrix. However in this case, it does not lead to closed-form solutions. Numerical optimization was very unstable in practice, therefore we consider the parameters to be fixed, as in the MRF and CRF cases.

Update of the parameters of the local lower bounds

The tightness of the local bounds depends on additional parameters which can be optimized following the same approach as for the prior parameters [Bishop, 2006]. The JJ bound and the Böhning bound introduced the parameters ξ, while the bound over the square root function for the TV prior introduced the additional parameter u. We get therefore $\xi_n^2 = \mathbb{E}[(\mathbf{\Phi}_n^T \mathbf{W})^2]$ for the GLSP prior, $\xi_n^2 = \mathbb{E}[w_n^2]$ for the FDSP and TV priors and $\xi = \mathbb{E}[\mathbf{W}]$ for the GP prior, noting $\xi = [\xi_1, \ldots, \xi_N]^T$. Finally, $u_n = \sum_d \mathbb{E}[(w_n - w_{\delta_{d(n)}})^2]$.

3.4.4 An incremental and sparse algorithm for the GLSP prior

In section 3.3.4, we established a connection between the GLSP prior and the Relevance Vector Machine (RVM) framework. The RVM was introduced for regression or binary classification and is an example of sparse Bayesian learning. The RVM also defines a zero mean Gaussian prior over $\mathbf{W}$ with a diagonal covariance matrix. However, in this case, each weight is associated with its own precision parameter α_l, instead of sharing the same α as in Eq. 3.15, for $\mathbf{R} = \mathbf{I}_L$. In practice during inference, many of the α_l tend to infinity meaning that the associated basis functions are irrelevant and can be removed from the model [Tipping, 2001]. Therefore adopting a diagonal instead of a spherical covariance leads to a selection of the most relevant basis. However, initializing with the full set of basis functions as proposed by the first RVM algorithm leads to a high computational cost due to the matrix inversion steps.

After a careful analysis of the log marginal likelihood, Tipping and Faul suggested a far more efficient algorithm [Tipping & Faul, 2003] which constructs the set of relevant basis functions in an incremental manner. It is initialized with a small number of active basis functions, and then adds or removes basis functions at each iteration, selecting the action leading to the largest increase in log marginal likelihood. This strategy enables keeping a reasonable number of relevant basis functions at all times, leading to faster computations and sparser solutions [Tipping & Faul, 2003]. In [Sabuncu & Van Leemput, 2012], the first greedy RVM algorithm was extended to the context of image-based prediction and image classification. In this section, we generalize the more efficient framework of Tipping and Faul to the domain of image segmentation, which has, to the best of our knowledge, never been proposed before.

The level of regularization depends on some characteristics of the basis functions, such as the spacing between their centers or their radii. In the variational approach presented earlier, all basis functions are present in the model and no selection is performed, leaving the decision to the user, for example guided by grid-search. The rationale behind the sparse extension is thus to allow a better, automatic and data-driven, selection of the relevant basis functions.

To this end, we no longer wish to maximize the data log likelihood (Eq. 3.16), but the following log joint probability:

$$\begin{aligned} \log p(I, \mathbf{W}|\alpha) &= \log p(I|\mathbf{W}) + \log p(\mathbf{W}|\alpha)\,, \\ &\geq \mathcal{L}(q) + \log p(\mathbf{W}|\alpha)\,, \end{aligned} \tag{3.42}$$

where $\alpha = [\alpha_1, \ldots, \alpha_L]^T$ and the lower bound $\mathcal{L}(q)$ is now written:

$$\mathcal{L}(q) = \sum_Z \sum_\Gamma \int_{\theta_I} q(Z)q(\Gamma)q(\theta_I) \log \frac{p(I, Z, \Gamma, \theta_I|\mathbf{W})}{q(Z)q(\Gamma)q(\theta_I)} d\theta_I \,. \tag{3.43}$$

Instead of relying on variational calculus, the weight posterior approximation $q(\mathbf{W})$ is computed using a Laplace approximation, corresponding to a second-order Taylor expansion of the lower bound around the mode $\mu_\mathbf{W}$ maximizing Eq. 3.42:

$$\log q(\mathbf{W}) = \log \mathcal{L}(q, \mu_\mathbf{W}) + \log p(\mu_\mathbf{W}|\alpha) - \frac{1}{2}(\mathbf{W} - \mu_\mathbf{W})^T \mathbf{\Sigma}_\mathbf{W}^{-1} (\mathbf{W} - \mu_\mathbf{W}) + \text{cst} \,. \tag{3.44}$$

The mode $\mu_\mathbf{W}$ and the Hessian matrix at the mode are found through a Gauss-Newton optimization formulated as an iterative reweighted least squares (IRLS) algorithm. This leads to the following expression for the covariance:

$$\mathbf{\Sigma}_\mathbf{W} = [\mathbf{\Phi B \Phi}^T + \mathbf{A}]^{-1} \,, \tag{3.45}$$

where $\mathbf{B} = \text{diag}([-g_1''(\mathbf{\Phi}_1^T\mathbf{W}), \ldots, -g_N''(\mathbf{\Phi}_N^T\mathbf{W})])$ is a diagonal matrix, g_n'' being the second derivative of the function defined as $g_n(x) = \eta_{n1} \log \sigma(x) + (1 - \eta_{n1}) \log \sigma(-x)$. Furthermore, $\mathbf{A} = \text{diag}(\alpha)$.

The precision parameters α are updated by following a type-II maximum likelihood approach, which corresponds to maximizing Eq. 3.43 after marginalization of the variable weights, giving: $\mathcal{L}(\alpha) = \int_\mathbf{W} \mathcal{L}(q) + \log p(\mathbf{W}|\alpha) d\mathbf{W}$. Introducing the matrix $\mathbf{C} = \mathbf{B}^{-1} + \mathbf{\Phi}^T \mathbf{A} \mathbf{\Phi}$ and $\hat{\mathbf{t}} = \mathbf{\Phi}^T \mu_\mathbf{W} + \mathbf{B}^{-1} g'$ where $g' = [g_1'(\mathbf{\Phi}_1^T\mathbf{W}), \ldots, g_N'(\mathbf{\Phi}_N^T\mathbf{W})]^T$, we can compute the derivative $\frac{\partial \mathcal{L}(\alpha)}{\partial \alpha_l}$ in closed-form as a function of α_l, of $\mathbf{C}_{-l}$ computed with the already selected basis and of $\hat{\mathbf{t}}$. We can use this relation to evaluate the gain in lower bound associated with each basis function and select the one leading to the largest gain [Tipping & Faul, 2003]. As for the RVM, three actions are then possible with respect to the candidate: if the basis function is already in the model, it can be either removed or its precision parameter re-estimated. If the basis function is not yet in the model, it is added to the set of active basis functions. A more detailed derivation of the incremental algorithm can be found in appendix A.4.

The resulting algorithm is efficient since it is constructive and causes a selection of a small subset of basis functions from the initial user-defined dictionary.

3.5 Results

3.5.1 Implementation of the algorithm

The proposed segmentation algorithm can be used to segment a whole image and gather voxels according to their appearance and localization. When the spatial smoothness is induced by an MRF prior, it is then a simple extension of the neighborhood EM algorithm introduced in [Ambroise et al., 1997], with mixtures of Gaussians instead of a unique component to model the appearance of each region, and with a full variational inference approach allowing posterior distributions over the intensity parameters to be estimated.

Furthermore, we also consider the case where an initial prior probability is provided. This corresponds to the situation for instance where the user knows where the structure

to be segmented is located in space. The user provides a subset of marked voxels acting as foreground seeds. The label prior for these voxels is fixed to 1. Elsewhere, it is initialized to 0, except on a narrow band of width r defined by the user along the foreground region. Here, the label prior goes from 1 for the voxels along the foreground to 0 for those along the background region.

Algorithm 1: Bayesian image segmentation

Initialization:
- *Initialize label prior* $p(Z)$
- *Initialize appearance parameters* θ_I

while *not converged* **do**
- **E-Step:** *Compute* $q(Z)$ *following Eq. 3.20*
- **MI-Step:** *Update* $q(\Gamma)$ *and* $q(\theta_I)$ *from Eq. 3.21 and 3.22 to 3.25*
- **MP-Step:**
  - **if** *narrow band* **then**
    - *Update the narrow band from* $q(Z)$
  - **end**
  - *Update* $q(\mathbf{W})$, $q(\xi)$ *and prior hyperparameters*

end

The label posterior $q(Z_n)$ is then trivial for the regions where $p(Z_n) \in \{0, 1\}$. It only needs to be computed for voxels on the narrow band during the E-step. Before updating the posterior approximations for the regularization variables, a new narrow band $\mathrm{NB}_{\mathrm{new}}$ of width r is defined along the isoprobability surface 0.5 of the label posterior $q(Z)$. Spatial regularization is then performed only on the narrow band during the MP-step, while the label prior for voxels outside of the narrow band remains unchanged. These steps are summarized in Fig. 3.5.

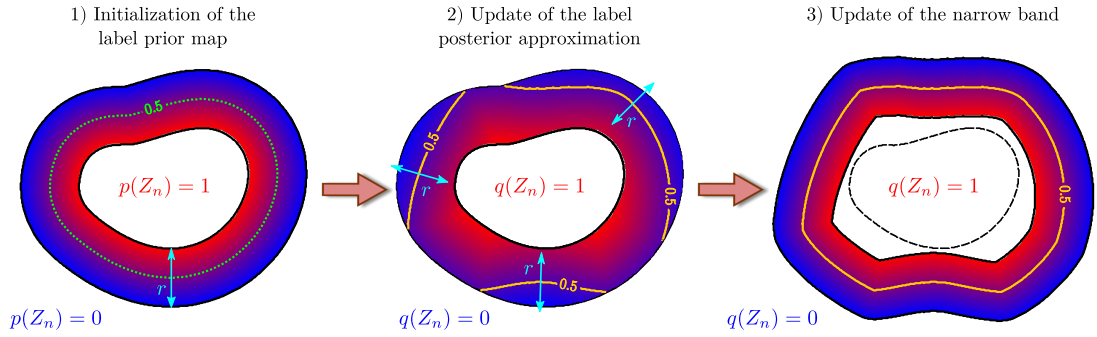


Fig. 3.5: Update of the narrow band for the algorithm initialized with a user prior.

In this way, the foreground region will grow or shrink at each iteration until convergence, reached when the narrow band stops evolving. This approach is similar to the fast level set method proposed by [Adalsteinsson & Sethian, 1995], where only the voxels close to the zero-level set are used for the computations. The sketch of the two approaches (whole-image and narrow-band-based evolution) are summarized in Alg. 1.

Finally, the algorithm can also be used in a supervised way. The intensity parameters are given and the objective is to regularize the segmentation in space. The

variational approach leads to an iterative estimate of the label posterior $q(Z)$ and the prior parameters.

3.5.2 Whole image vs narrow band evolution

In this section we provide a visual comparison between the whole-image approach and the algorithm based on an evolving narrow band for two segmentation cases, focusing either on a lung nodule or on a kidney. The first case was extracted from the LIDC dataset [Armato III et al., 2011], which is a publicly available database of pulmonary CT scans with lung nodule annotations. The kidney CT image was obtained from the dataset of the QUBIQ challenge [Menze et al., 2020].

The two examples are presented in Fig 3.6 and Fig. 3.7, respectively. The first one is regularized with a TV prior while the second is using a GLSP prior. The contour provided by the user to initialize the narrow band algorithm is shown in the upper right image for both cases. The whole-image approach is initialized using the K-means algorithm.

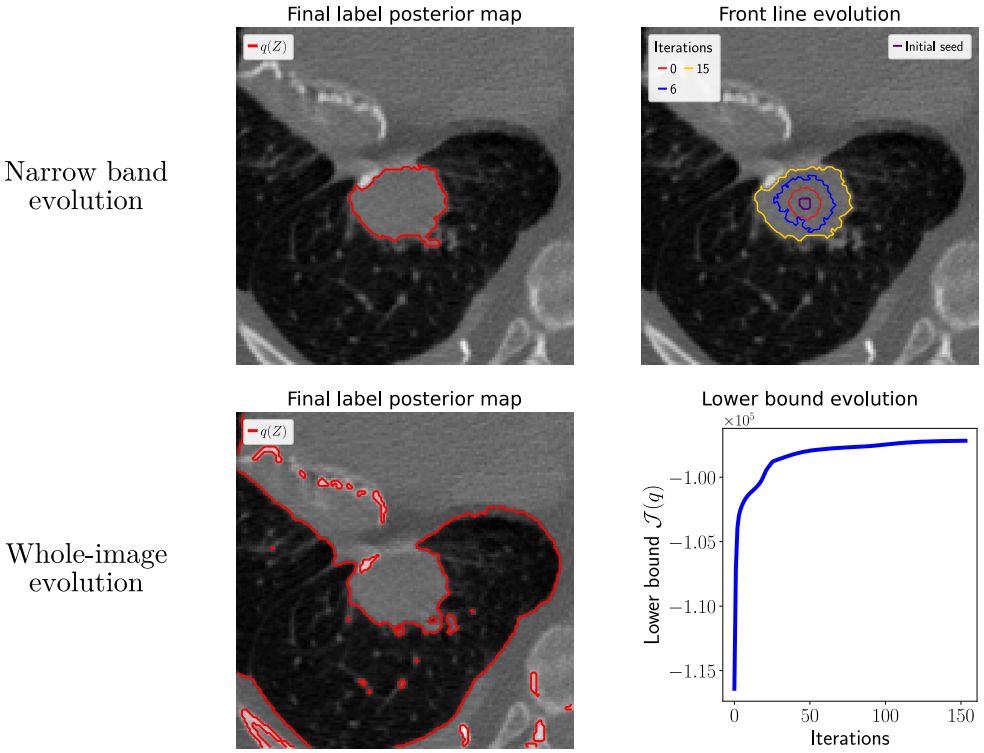


Fig. 3.6: Comparison between the narrow band algorithm and the whole-image approach on an image from the LIDC dataset. Both are fitted with a TV prior.

Results show that the narrow band algorithm is better suited for segmenting a particular structure in the image. Indeed, if the appearance models of the two regions are sufficiently different, the narrow band stops its evolution after reaching the structure boundaries, thus producing a proper segmentation. On the contrary, the whole-image approach produces a segmentation leaking outside of the object if the image contains other regions with intensity patterns similar to the structure of interest. For instance, the segmentation produced by the whole-image approach leaks outside the lung nodule on

Fig. 3.6 and outside of the kidney on Fig. 3.7. Appearance models of the foreground and background regions learnt by the narrow band algorithm for the kidney segmentation are shown in Fig. 3.8.

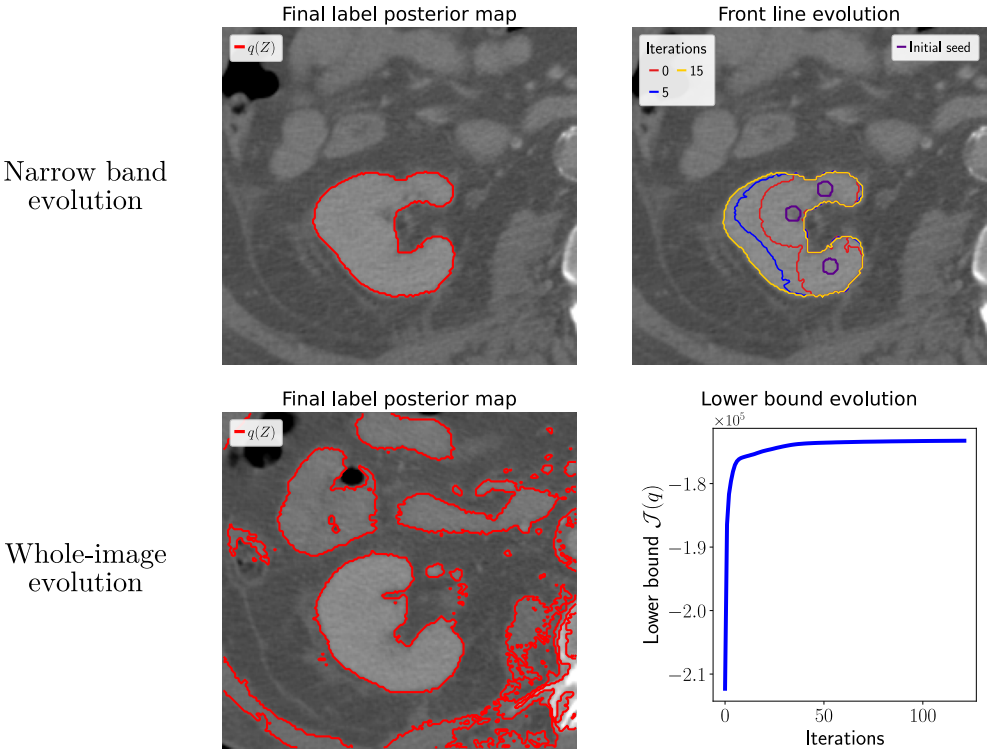


Fig. 3.7: Comparison between the narrow band algorithm and the whole-image approach on an image from the LIDC dataset. Both are fitted with a GLSP prior.

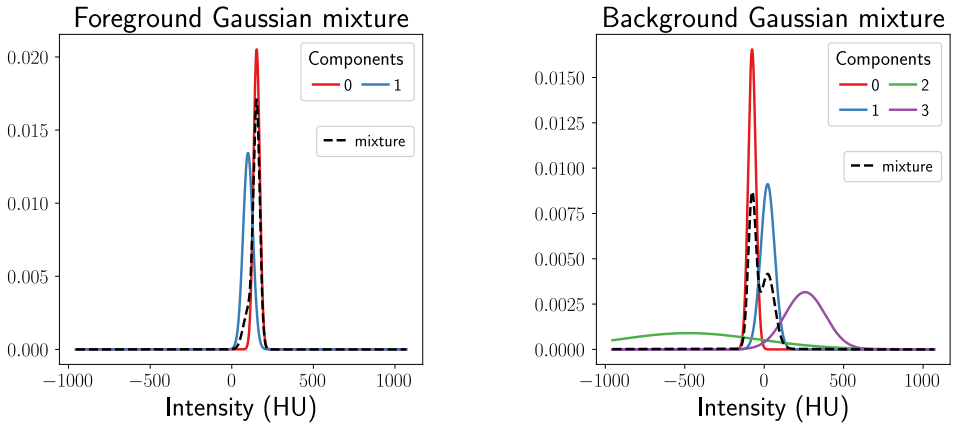
The lower bound $\mathcal{J}(q)$ can be computed for the whole-image approach and is used as a stopping criterion. Details about the computation are given in appendix A.2. Its evolution is shown on the two figures for both cases. As it cannot be computed with the narrow band approach, the algorithm is stopped in this case when the narrow band stops evolving.

3.5.3 Spatial priors comparison

In this section, we analyse the influence of the priors over the segmentation smoothness and the effect of the prior hyperparameters. The comparison is performed on the kidney CT scan from the QUBIQ dataset. A whole-image approach is used, with one mixture component in each region.

Fig. 3.9 presents the results for the six spatial priors: the MRF, CRF, TV, FDSP, GP (with two optimization strategies) and GLSP priors. Each column corresponds to one parameter setting defined by the user. The 0.5 isoprobability contour of the label prior is shown next to the one of the label posterior approximation, when possible.

We use a 4-connectivity neighborhood for the MRF and CRF priors in both parameter settings, which differ with respect to the value of the parameter β. For $\beta = 0.4$, the segmentation with a CRF prior exhibits smoother contours than the one obtained with an MRF. The increase in β mainly has an impact on the MRF segmentation, with a



(a) **(b)**

Fig. 3.8: Gaussian mixtures of the foreground (3.8a) and background regions (3.8b) of the kidney segmentation by the narrow band algorithm.

stronger regularization then leading to smoother contours, comparable to those of the CRF. The difference between the two priors is explained by the fact that the MRF is purely based on connectivity and does not take contrast information into account.

The TV and FDSP priors lead to similar posterior maps. No parameter needs to be set by the user, thus the second column is left blank. Unlike the MRF and CRF priors, the hyperparameter α is adjusted automatically during the inference. Given the influence that β has on the result's smoothness, it is a great advantage here to be able to learn the appropriate level of regularization from the data.

The next two rows are results with a GP prior but with two different optimization approaches. The length scale ω_1 is increased from 2 to 5 between the two columns leading again to a stronger regularization. The hypothesis of periodic boundary conditions for the Fourier optimization strategy has no impact here as the length-scale remains small in comparison with the image size. The two strategies lead, therefore, to very similar results. This may be different, however, for smaller images or larger length scales; in these cases the exact approach based on Kronecker products is preferable. The lower bound is tractable here (Fig. 3.10), making it possible to implement a grid search approach to set the parameters.

Finally, the last row shows results with a GLSP prior. The matrix $\mathbf{R}$ is chosen to be the identity $\mathbf{I}_L$. The two parameter settings correspond to two different spacings, s, between the basis function centers and two different radii, r, larger values giving smoother results. Though the hyperparameter α is tuned during the inference, the user still needs to provide the dictionary of basis functions. One solution is to perform a grid search as for the GP prior, or to use the incremental algorithm. A comparison between the greedy and incremental approaches will be presented in section 3.5.5.

An attractive property of variational inference methods is the possibility to monitor the convergence of the model by following the evolution of the lower bound. Moreover, the lower bound can also be used to perform model selection by comparing the values reached after convergence. However, this implies to be able to take all constant terms into account, which is not possible for the TV, MRF and CRF priors due to the intractability

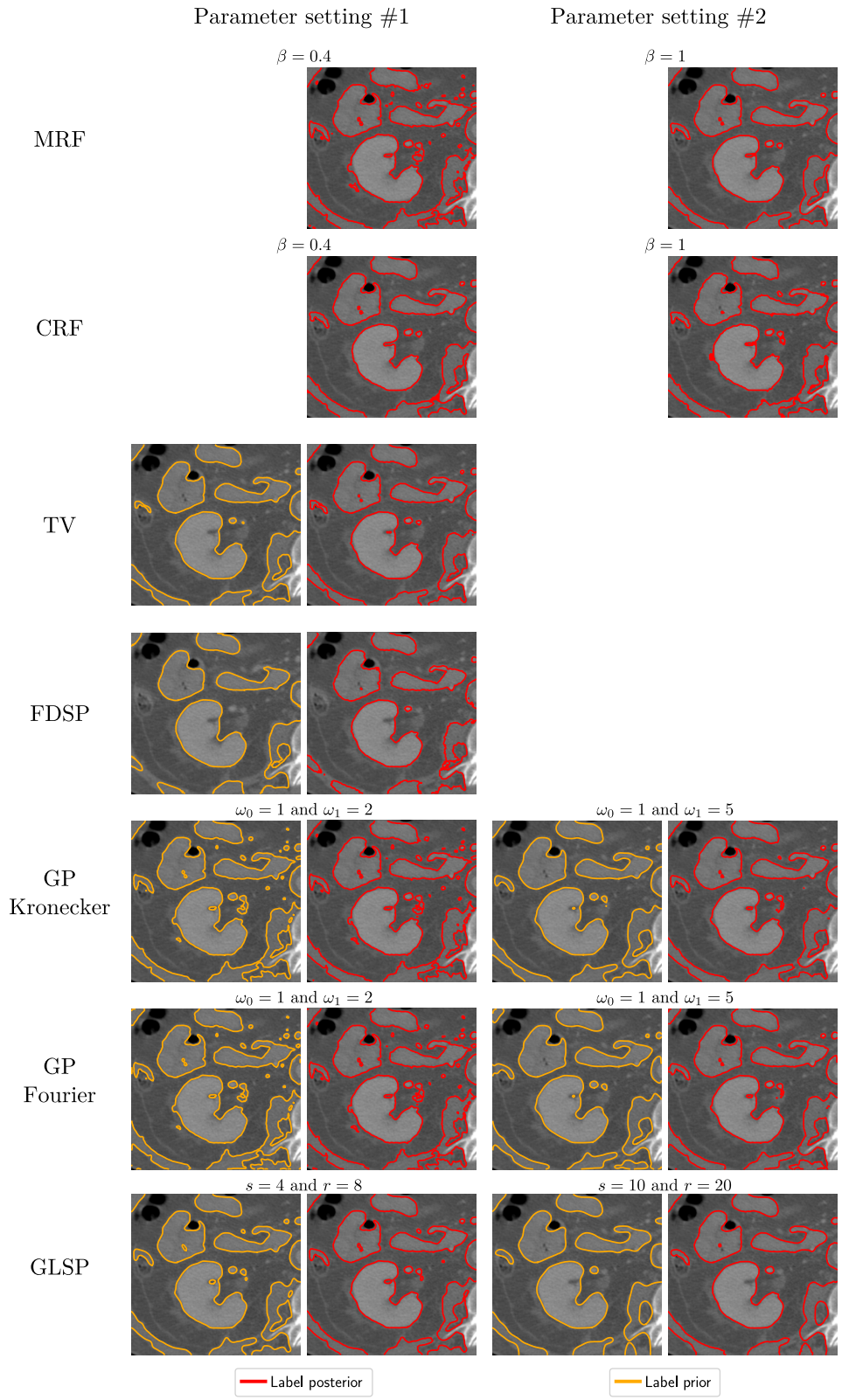


Fig. 3.9: Visual comparison of the different spatial priors with two parameter settings. The label posterior isoline $q(Z) = 0.5$ is shown in red, together with the label prior contour $p(Z) = 0.5$ in yellow when possible.

of their normalization factors. For the others, the constants are known and a comparison can be made. Fig. 3.10 compares the values reached by the lower bound at convergence for the FDSP, GP and GLSP priors. We can observe that the GLSP prior leads to higher lower bound values than the GP prior. In section 3.3.4, we discussed how the two are related. By defining basis functions, the GLSP is performing a subsampling of the image grid, leading to a sparser model than the GP, which is working with a full covariance matrix. The difference in lower bound could then be explained by the Ockham's razor principle, stating that the less complex models should be favoured.

Regarding the computation time, results are presented in Fig. 3.11. Convergence is assumed when the increase in lower bound between 2 iterations falls below a defined threshold. The GLSP is the most time-consuming prior and the inference time depends heavily on the number of basis functions inside the model (2400 and 380 for the first and second setting, respectively). The other priors have fast and comparable computational times. For larger images, the GP prior optimized with the Fourier approach will be faster than the one based on Kronecker products.

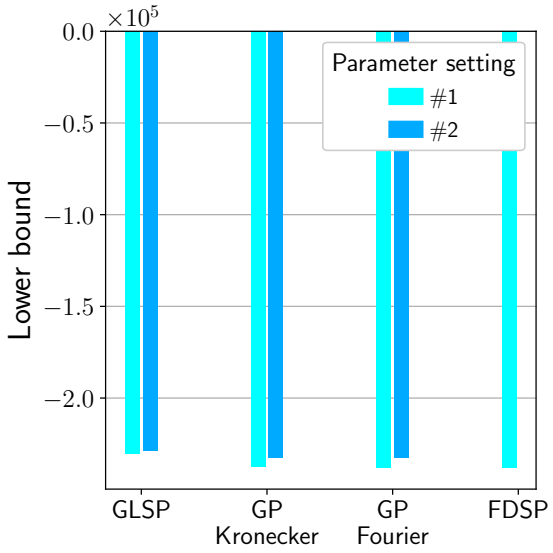


Fig. 3.10: Lower bound values after convergence for several spatial priors and two parameters settings.

3.5.4 Uncertainty quantification

Being able to assess the uncertainty of a segmentation is of particular interest, especially for medical applications. With a generative model, one solution is to produce new segmentation samples allowing the regions with higher variability to be identified.

The mean field variational inference learns a label posterior approximation that factorizes over the pixels in the image: $q(Z) = \prod_n q(Z_n)$. The simplest approach to generate new segmentation maps is to sample the Bernoulli distributions $q(Z_n)$ independently for each pixel. MCMC approaches like Gibbs sampling produce more accurate samples by taking into account the spatial interactions between pixels [Morris et al., 1997]. However, these methods also have their limitations, because they require a large number of iterations to be able to take into account long range correlations

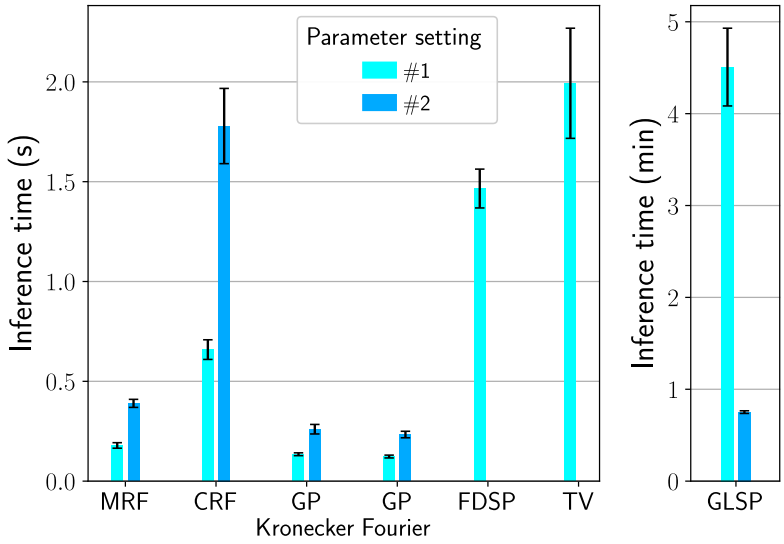
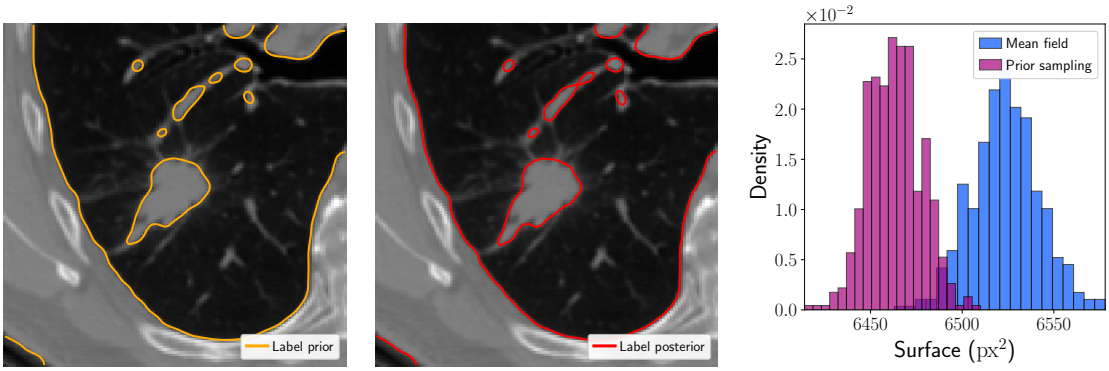


Fig. 3.11: Computation times for several spatial priors and two parameter settings.

[Stoehr, 2017]. On the other hand, a third solution exists for the GLSP and GP priors as the posterior approximation does not factorize over the vector $\mathbf{W}$. It allows new prior samples to be generated that take into account the spatial correlations between pixels. Posterior samples are then obtained by applying Bayes' theorem.



(a) **(b)** **(c)**

Fig. 3.12: Uncertainty quantification. (3.12b) Label prior contour. (3.12c). Label posterior contour. (3.12a) Area distribution over 500 samples.

In this section, we compare the simplest sampling approach that assumes independence between pixels with the Bayesian approach on an image extracted from the LIDC dataset. The model is fitted with a GLSP prior. In order to focus on the effect of the prior, the comparison is performed within a supervised framework with fixed intensity parameters. The label prior and posterior contours are shown in Fig. 3.12a and Fig. 3.12b, respectively. Four samples obtained by sampling the Bernoulli distributions are shown in Fig. 3.13. The irregular contours and numerous isolated pixels show that the independence hypothesis is a limitation. Conversely, first sampling $q(\mathbf{W})$ and then computing the posterior through Bayes' rule leads to smoother and more plausible samples as shown in Fig. 3.13.

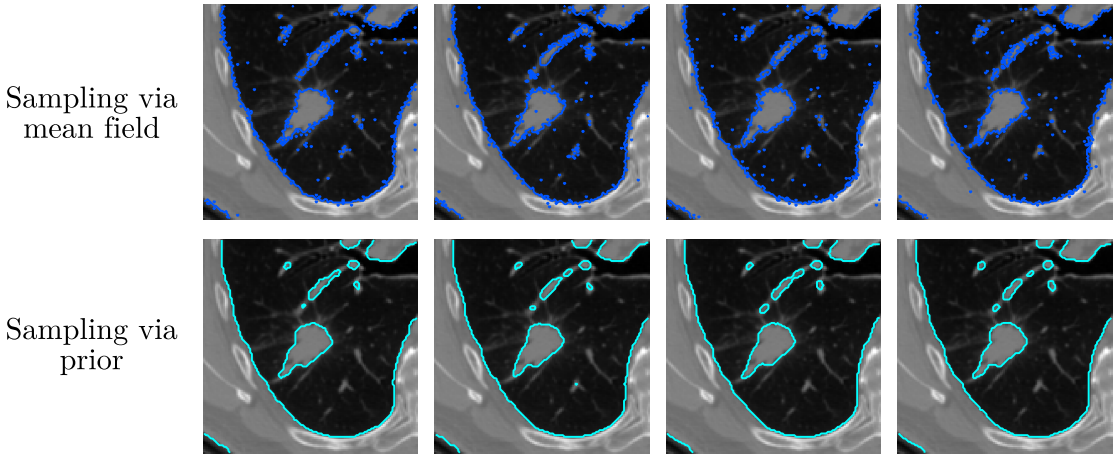


Fig. 3.13: Samples generated by sampling the Bernoulli distributions are less realistic than those obtained after sampling the prior distribution.

3.5.5 GLSP incremental algorithm

Finally, we compare the greedy GLSP algorithm with the incremental one. All results presented before with the GLSP prior were produced using only one basis function setting at each time, i.e., only one spacing s and only one radius r. Yet, we demonstrated in section 3.5.3 the influence of these parameters over the level of regularization. To find the appropriate combination, one can perform a grid search and select the values maximizing the lower bound at convergence. A more elegant approach is to use the incremental algorithm introduced in section 3.4.4 which performs an automatic, data-driven, selection of the basis functions maximizing the lower bound.

An image from the LIDC dataset is once again used to perform the comparison, and we define two categories of basis functions differing in their spacing and radius values, $\{s = 10$ px, $r = 20$ px$\}$ or $\{s = 20$ px, $r = 50$ px$\}$, respectively. The basis functions are spread over a regular grid in the image, leading to a total of 218 basis functions. The image, with a visualization of the radii, is presented in Fig. 3.14.

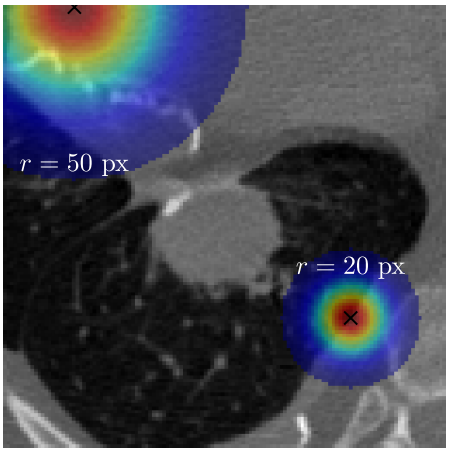
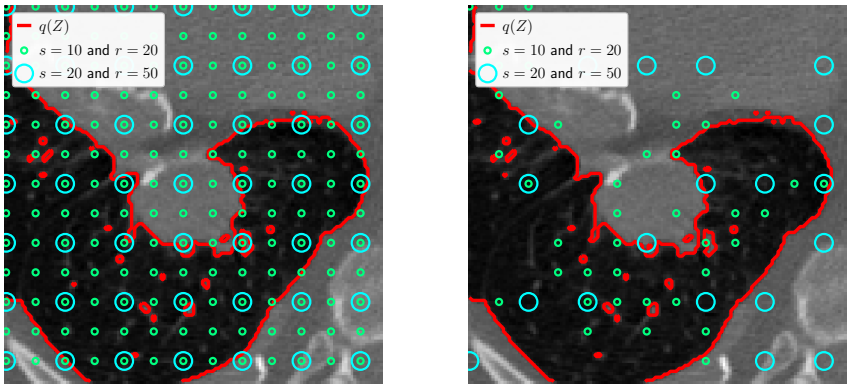


Fig. 3.14: Visualization of the radii r of the basis functions used for comparison between the incremental and the greedy GLSP algorithm.

Label posterior 0.5 contours for both algorithms fitted on the whole image are shown in Fig. 3.15. The first observation is that the two contours are very close. In the greedy algorithm, all basis functions are active, meaning that $L = 218$ all the time. In contrast, the incremental algorithm performs a selection, as can be seen in Fig. 3.15b. At convergence, there are 58 relevant basis functions, i.e., 27% of the total dictionary. We can see that basis functions with large radii are selected in wide areas of uniform intensity contrast while smaller radii are preferred along the posterior boundaries.



(a) **(b)**

Fig. 3.15: Comparison between the greedy (3.15a) and incremental (3.15b) algorithms using the GLSP prior. The circles indicate the position of the centers of the active basis functions after convergence.

Regarding the computation time, the greedy algorithm is faster than the incremental one for this particular case. The main burden for the greedy algorithm is the inversion of the covariance matrix defined in Eq. 3.31. Here, the total number of basis functions is not too high, leading to a reasonable computation cost. On the contrary, several steps in the IRLS procedure are required to approach the mode of the Laplace approximation, penalizing the incremental algorithm.

Finally, we analyse the effect of changing the regularizer encoded in the matrix $\mathbf{R}$ when using the greedy approach. Previously, we were penalizing the Euclidean squared norm of the vector $\mathbf{W}$ by setting $\mathbf{R} = \mathbf{I}_L$. Instead, we can choose to penalize the magnitude of the derivatives of the prior label field $||D^i f||^2$, where D is a linear differential operator.

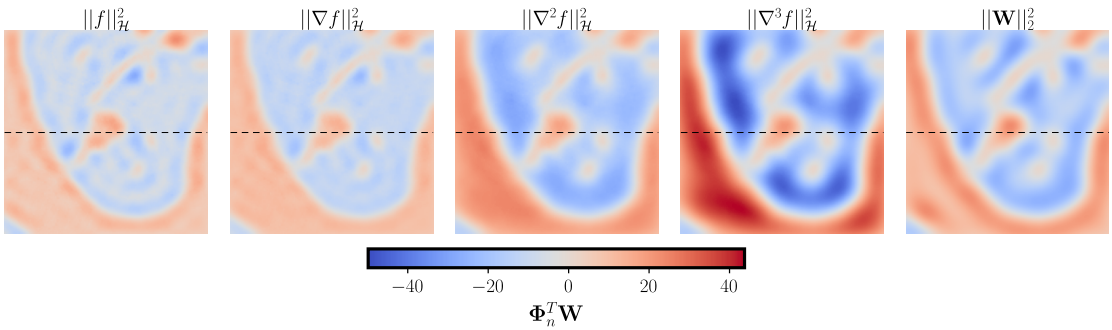


Fig. 3.16: Prior label field obtained with several regularizers. The evolution of the field along the dashed line is plotted in Fig. 3.17.

Fig. 3.16 provides a visual comparison between the label prior fields $f = \mathbf{\Phi}^T\mathbf{W}$ obtained for $i = 0, 1, 2$ or 3, on an LIDC image. Again, we use a supervised setting to focus on the regularization part. We can observe that we get smoother results with higher derivatives. The field also takes larger absolute values which will push the prior $p(Z)$ closer to 0 or 1 after applying the sigmoid. The evolution of the field along the dashed line is shown in Fig. 3.17. One can observe that penalizing the Euclidean squared norm of the weights vector already provides good smoothing properties, similar to penalizing the second derivatives.

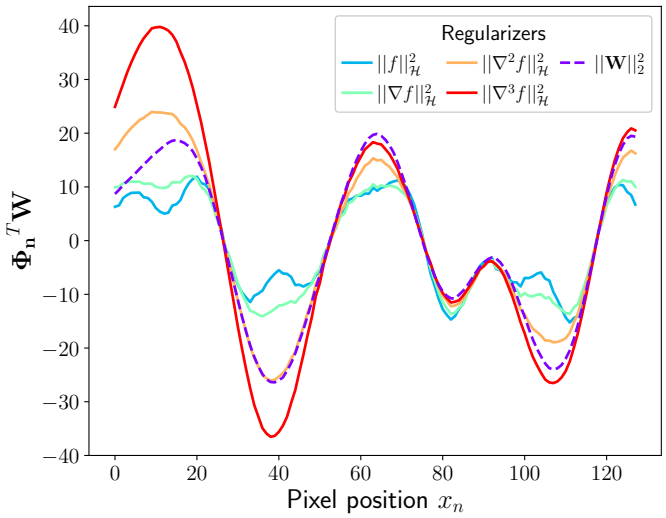


Fig. 3.17: Prior label field evolution along a row of the image for several regularizers.

Moreover, changing the regularizer does not affect the main boundaries in the image. Instead, it mostly changes the slope of the transition between regions, which has in practice little effect on the label posterior after applying the sigmoid and combining with the intensity likelihood.

3.6 Conclusion

Image segmentation is a key pre-processing task in computer vision. A natural Bayesian approach to image segmentation is to combine image-derived information with some prior knowledge regarding the result. In this chapter, we focused on priors enforcing spatial consistency, in order to be able to capture the spatial correlations that exist intrinsically between pixels of an image. Four families of priors were reviewed, including the classical MRF and a prior based on a GP formulation, which has, to the best of our knowledge, never been proposed before.

We suggested a common inference scheme based on variational calculus and local variational bounds allowing the prior formulation to be conveniently changed. The framework can be used as a standalone segmentation method, either to perform spatial clustering on a whole image or to segment a particular structure. For the latter case, we introduced a method based on the evolution of a narrow band, starting from a user input. The scheme can also be used in a supervised fashion, to regularize a segmentation when the appearance models are given.

We provided a visual comparison between the different priors and showed the influence of their hyperparameters on the regularization level. Although the priors performed in a similar way, three of them, the TV, FDSP and GLSP priors, have the advantage of having their hyperparameter α being adjusted automatically by the model. The TV prior does not have any parameter requiring to be fixed by the user. However, this is explained by the fact that the size of the neighborhood taken into account at each pixel location is fixed, by definition of the total variations. In contrast, for the FDSP prior for example, it depends on the derivative order that the user chooses to consider. The incremental version of the algorithm for the GLSP prior enables somehow to let some parameters be chosen by the model and allows the regularization level to be adapted locally, but it remains computationally slower than the greedy formulation.

If the user wants to enforce a high level of regularization in the segmentation, the TV and FDSP priors are not the most suitable priors as they consider by definition a small neighborhood around each pixel. To take into account longer-range correlation, a solution for the FDSP is to penalize higher derivatives. For the other priors, such correlation can be achieved by simply increasing the neighborhood radius of the MRF or CRF priors, the length scale of the GP or the set spacing/radius for the GLSP.

Proper quantification of the segmentation uncertainty is enabled by the generative nature of the model. In particular, the GLSP and GP priors allow efficient sampling of the label posterior without resorting to slower MCMC approaches while taking into account the spatial correlations between pixels.

The main characteristics of each prior are summarized in Tab. 3.1.

Tab. 3.1: Comparison summary of the six spatial priors.

	MRF	CRF	TV	FDSP	GP	GLSP
Computation time	++	++	++	++	+	+++
# of user-fixed parameters	2	3	0	1	2	2
Tractable lower bound	✗	✗	✗	✓	✓	✓
Regularization power	+++	++	+	++	+++	+++
Non-diagonal covariance matrix	✗	✗	✗	✗	✓	✓
Locally adaptive regularization	✗	✗	✗	✗	✗	✓

Only binary segmentation problems were considered in this chapter, but working with several classes may be interesting for a whole-image approach. The extension to the multi-class setting is straightforward, and involves replacing the sigmoid link function by the softmax function and using the Böhning bound in all cases. In addition, the problem of choosing the right number of classes can be addressed with a Dirichlet process which allows the appropriate number of clusters to be automatically selected [Orbanz & Buhmann, 2008].

Furthermore, one must note that the list of spatial priors detailed in this chapter is not exhaustive. Other formulations exist, for instance based on wavelet decomposition [Figueiredo, 2005a].

Finally, an interesting future perspective would be to consider nonstationary covariance functions for the GP prior. The GLSP incremental algorithm already allows the level of regularization to be adapted depending on the location in the image, by placing basis functions with large scales in uniform areas and smaller scales close to the

transitions. Allowing the GP prior to be locally adaptive while preserving the scalability is an open challenge for future work.

Acknowledgments

This work was partially funded by the French government, through the UCA^{JEDI} and 3IA Côte d'Azur "Investments in the Future" projects managed by the National Research Agency (ANR) with the reference numbers ANR-15-IDEX-01 and ANR-19-P3IA-0002 and supported by the Inria Sophia Antipolis - Méditerranée "NEF" computation cluster.

Chapter 4

Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model

Contents

4.1 Introduction . . . 46
4.2 Unsupervised quality control workflow . . . 49
4.2.1 Input segmentation . . . 49
4.2.2 Probabilistic model . . . 50
4.2.3 Detection of challenging cases . . . 50
4.2.4 Use case . . . 51
4.3 Method . . . 51
4.3.1 Mixtures of multivariate Student's t-distributions . . . 51
4.3.2 Spatial smoothness prior . . . 51
4.3.3 Implementation . . . 53
4.4 Probabilistic inference . . . 55
4.4.1 MRF regularization . . . 55
4.4.2 GLSP regularization . . . 55
4.4.3 FDSP regularization . . . 55
4.5 Results . . . 57
4.5.1 Datasets . . . 57
4.5.2 Unsupervised indices . . . 58
4.5.3 Setting hyperparameters . . . 58
4.5.4 Qualitative analysis . . . 60
4.5.5 Quantitative analysis . . . 63
4.5.6 Results interpretability . . . 67
4.5.7 Surrogate segmentation performance . . . 69
4.5.8 Discussion . . . 70
4.6 Conclusion . . . 70

Monitoring the quality of image segmentation is key to many clinical applications. This quality assessment can be carried out by a human expert when the number of cases is limited. However, it becomes onerous when dealing with large image databases, so partial

automation of this process is preferable. Previous works have proposed both supervised and unsupervised methods for the automated control of image segmentations. The former assume the availability of a subset of trusted segmented images on which supervised learning is performed, while the latter does not. In this chapter, we introduce a novel unsupervised approach for quality assessment of segmented images based on a generic probabilistic model. Quality estimates are produced by comparing each segmentation with the output of a probabilistic segmentation model that relies on intensity and smoothness assumptions. Ranking cases with respect to these two assumptions allows the most challenging cases in a dataset to be detected. Furthermore, unlike prior work, our approach enables possible segmentation errors to be localized within an image. The proposed generic probabilistic segmentation method combines intensity mixture distributions with spatial regularization prior models whose parameters are estimated with variational Bayesian techniques. We introduce a novel smoothness prior based on the penalization of the derivatives of label maps which allows an automatic estimation of its hyperparameter in a fully data-driven way. Extensive evaluation of quality control on medical and COCO datasets is conducted, showing the ability to isolate atypical segmentations automatically and to predict, in some cases, the performance of segmentation algorithms.

This chapter corresponds to the following publications:

- [Audelan & Delingette, 2021] **B. Audelan** and H. Delingette. Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model. *In Medical Image Analysis, 68, 2021, p. 101895.*
- [Audelan & Delingette, 2019] **B. Audelan** and H. Delingette. Unsupervised Quality Control of Image Segmentation Based on Bayesian Learning. *In Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pp. 21–29.*

4.1 Introduction

Semantic segmentation of an image is the process of associating a label to every pixel in an image. This task is particularly important in a medical context since it impacts downstream algorithms using image segmentations as input, but also the decisions that clinicians may make about the patient. For instance, in radiotherapy planning, the delineations of tumor lesions directly influence the extent of the dose delivered around the tumor. Also, obtaining reliable image segmentations is mandatory to use image derived biomarkers in a clinical setting [Keshavan et al., 2018]. Finally, the development of supervised learning for image segmentation requires the accumulation of potentially large sets of manually or semi-manually segmented image databases that need to be quality controlled. Such segmentations are prone to inter-rater variability [Visser et al., 2019] in addition to plain errors. It is therefore of great importance to automatically detect possible failed segmentation cases, whether those segmentations are generated by an algorithm or a human rater. The challenge is to perform this monitoring in the absence of ground truth segmentations.

In prior work, evaluation methods can be categorized either as supervised or as unsupervised, depending on whether a reference segmentation is required or not [Zhang et al., 2008]. A first set of supervised methods is based on a classifier which accepts or rejects the proposed segmentation based on combined features. For instance, in

[Hui Zhang et al., 2006], decision trees based on handcrafted features depending on the image (texture, color space...) and on the geometry of the segmented region (perimeter, compactness...) are combined in a single classifier.

In [Shamir & Bomzon, 2019], a decision tree predicts the Dice score of head segmentations with an application to the treatment of brain tumors. In [Xu et al., 2009], a framework to detect failures in cardiac segmentation based on a shape parameter and an intensity feature has been proposed. The number of features taken into account is increased in [Kohlberger et al., 2012], where the model decision relies on 42 shape and appearance features. They are combined in an SVM classifier regressing the Dice coefficient between the given segmentation and the unknown ground truth. While in [Xu et al., 2009] the features were specific to cardiac segmentation, the approach taken in [Kohlberger et al., 2012] is more generic and was trained on segmentations of 8 different organs.

Reverse Classification Accuracy (RCA) has also been proposed for quality control assessment in [Valindria et al., 2017]. Assuming the availability of a set of trusted images with ground truth, the proposed segmentation on a new image is compared to the predicted one based on those reference images, which can result in rejection if discrepancies are too large. This approach was tested on larger databases in [Robinson et al., 2019] where the authors showed the ability of the method to highlight poor quality segmentations but pointed out the relatively long computation times as a bottleneck.

Another family of supervised approaches uses deep learning to estimate the quality of a segmentation. For instance, in [Robinson et al., 2018], a neural network is trained to predict the Dice coefficient of cardiac segmentations. The Jaccard index (intersection over union) is predicted by neural networks in [Arbelle et al., 2019; Huang et al., 2016; Shi et al., 2017] where the original image and the proposed segmentation mask are provided as input. Some authors have proposed exploiting the uncertainty of segmentations in order to assess their quality, within a deep learning framework. Uncertainty quantification also adds some interpretability to the quality assessment as it provides information about the location of possible errors. Bayesian QuickNat proposed by [Roy et al., 2019] uses Monte Carlo dropout at test time to generate several segmentation samples. The average over the samples gives the final segmentation map while variability across the different samples gives an estimate of the uncertainty of the segmentation. The authors show a good correlation between the measured uncertainty and the Dice coefficient between the segmentation and the unknown ground truth. Other methods to evaluate the uncertainty were explored in [Jungo & Reyes, 2019] and the results suggest that none is superior to the others. Finally in [DeVries & Taylor, 2018], a first network outputs a segmentation map and an uncertainty map at the pixel level, which are then taken as inputs by a second network which regresses a quality score at the image level.

A limitation shared by these methods is their supervised design, meaning that they require the extraction of a subset of segmented data that is considered to be "ground truth". This trusted subset is used by the models to learn how a "good" segmentation looks. The resulting decision rules making a new segmentation acceptable or not may thus be biased by the composition of the trusted set, which must be large enough for training a deep-learning-based framework. Further, access to large annotated datasets remains an issue in many domains including medical imaging. Finally, supervised methods often lack generality as their performance depends on the type of images and segmented structures in the training set.

In contrast, unsupervised approaches do not rely on a subset of trusted images but rather on assumptions about the appearance and shape of the foreground and background regions [Rosenberger et al., 2006; Zhang et al., 2008]. These assumptions are then translated into a set of segmentation metrics. For instance, common hypothesis is that a "good" segmentation exhibits high levels of intra-region homogeneity and inter-region heterogeneity [Johnson & Xie, 2011], and several handcrafted features have been proposed to measure them [Chabrier et al., 2006; Gao et al., 2017; Johnson & Xie, 2011; Zhang et al., 2008]. The main limitation of these approaches is that it is difficult to design discriminative indices and to find a proper way to combine them. Moreover, as mentioned by [Zhang et al., 2008], most of those metrics assume a single underlying intensity distribution, typically Gaussian, in both foreground and background regions which is overly simplistic and sensitive to outliers.

Last but not least, interpretability is a desirable property, as knowing the problematic regions could facilitate the segmentation curation. However, it is often an issue since many of the previous methods, supervised or not, are black boxes outputting a simple score, which does not help to understand why a segmentation has failed.

In this chapter, we propose a novel unsupervised approach for automated quality assessment of image segmentations. It is based on the comparison between a proposed segmentation S produced by an algorithm or a human rater and the segmentation M given by a generic probabilistic segmentation model. The generic model is based on two simple intensity and smoothness assumptions, the underlying hypothesis being that explainable segmentations correspond to clearly visible boundaries in the image well captured by M. On the contrary, segmentations far from M are categorized as difficult or challenging as they would require priors other than intensity and smoothness to be explained. The quality assessment of a set of segmented images is then performed by studying how the distance between the proposed segmentation S and the modelled segmentation M varies within the dataset. Segmentations that are lying on the tails of this distance distribution are considered to be atypical and are candidates for manual verification. We show the effectiveness of this approach to extract suspicious segmentations on various public datasets ranging from photographic images for object detection and segmentation (COCO dataset) to lung and brain medical images (LIDC and BRATS datasets). We also show that this approach can be used in some cases to predict the performance of segmentation algorithms.

Our main contributions are twofold:

- Instead of relying on an arbitrary subset of selected segmentations as a training set, we propose an unsupervised approach based on intensity and smoothness hypotheses without any prior knowledge of the structure to be segmented. It removes the bias related to the selection of the reference images and allows the quality of segmentations to be assessed when few or even no other segmentations are available from a database. Our method differs from previous unsupervised segmentation quality indices with a more complex and robust approach to modeling the intensity of the different regions in the image. In addition, it allows a combination of the key factors defining a "good" segmentation (i.e., the intra-region homogeneity and the inter-region heterogeneity) in a data-driven way. Last but not least, our method is visually interpretable. For instance, when dealing with 3D medical images, it allows automatic retrieval of the slices with suspicious segmentations. Finally, the result can be useful to guide the manual correction of poorly segmented cases.

- We provide different spatial regularization strategies to enforce the spatial continuity. In particular, we introduce a novel prior, denoted by FDSP (Finite Difference Spatial Prior), based on the penalization of the squared norm of the derivatives of the prior label map, which allows an adaptative learning of the hyperparameter. It is compared to the classical Markov random field (MRF) and another spatial prior based on a weighted combination of spatially smooth kernels introduced in an earlier work of the authors [Audelan & Delingette, 2019], which will be denoted by GLSP (Generalized Linear Spatial Prior) throughout the chapter.

This chapter expands [Audelan & Delingette, 2019] by proposing a different spatial regularization strategy for which the hyperparameter can be estimated. In addition, the novel regularization prior can be entirely inferred with a variational Bayes method (no Laplace approximation needed) and leads to much faster computations. We also provide more extensive experiments on different datasets and added a qualitative and quantitative comparison with unsupervised segmentation quality control indices proposed in prior works. The code with the different regularization strategies is available in this repository: `https://gitlab.inria.fr/epione/unsegqc`.

The rest of the chapter is organized as follows. Section 4.2 presents the general framework of our unsupervised quality control assessment. In section 4.3, we present our appearance model and the spatial priors. Section 4.4 describes the model inference depending on the regularization. Finally, we show in section 4.5 the relevance of our approach for segmentation quality control on several datasets.

4.2 Unsupervised quality control workflow

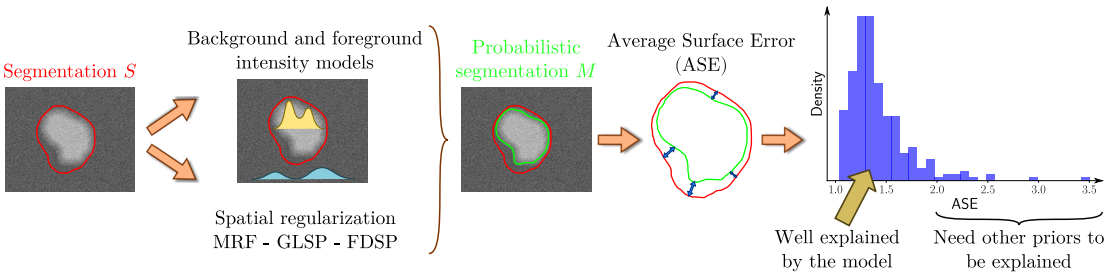


Fig. 4.1: Unsupervised segmentation quality control workflow.

Supervised segmentation quality control methods require the existence of a trusted subset of data from which quality assessment is learned. Instead, we follow an unsupervised approach (see Fig. 4.1) based on a probabilistic segmentation model relying only on two simple smoothness and intensity assumptions. Its great advantage is that it is agnostic with respect to the structure to be segmented and therefore can be run automatically even in the absence of ground truth.

4.2.1 Input segmentation

The input of the proposed method is a binary segmentation S on an image I into foreground and background regions for which we would like a quality estimate. There are no restrictions regarding the origin of S as it can have been created by an algorithm

or a human rater. Note that this is in contrast to several other methods that require the input segmentation to have been generated by a specific algorithm, like the uncertainty-based methods in deep learning [DeVries & Taylor, 2018; Jungo & Reyes, 2019; Roy et al., 2019].

4.2.2 Probabilistic model

Given the segmentation S, we produce a smooth contour or surface M close to S which is mostly aligned with visible contours in the image. We stress that the objective is not to build a surrogate ground truth, but instead to use M only as a comparison tool.

Intensity assumption. The first hypothesis of our approach is that intensity distribution variations in the image can help to understand segmentations. Given the segmentation S, two intensity models are built for the foreground and background regions.

Spatial smoothness assumption. The second hypothesis relies on the generally accepted assumption that two neighbouring voxels share a higher probability of belonging to the same label region. This is classically enforced by the use of discrete priors such as MRF. In [Audelan & Delingette, 2019], we proposed a regularization strategy based on a combination of spatially smooth kernels (GLSP). In addition to these two possibilities, we introduce in this chapter a novel way to take into account the spatial organization of the voxels, which we call the Finite Difference Spatial Prior (FDSP). This approach allows full tractability of the hyperparameter in an efficient manner which is not possible for the MRF and GLSP formulations.

The two assumptions are combined into a probabilistic model that outputs a new segmentation map M. By construction, M is typically a smooth contour which is mostly aligned with the visible intensity boundaries in the image. Again, M should only be seen as a representation used to benchmark the input segmentation S.

To measure the adequacy of S with respect to M, we employ the average asymmetric surface error (ASE) defined as $E_S = d(S, M) = \frac{1}{\partial S} \sum_{x \in \partial S} \min_{y \in \partial M} d(x, y)$ where ∂ denotes the segmentation surface. We discard the metric $d(M, S)$ as being uninformative since M is not a surrogate ground truth. An alternative measure to the ASE used in this chapter is the Dice score computed between the segmentations M and S.

4.2.3 Detection of challenging cases

Segmentations S close to M are identified as being explained by the model. In that case, the two intensity and spatial smoothness assumptions upon which the probabilistic model is based are sufficient to understand the contours. However, segmentations S far from M are classified as unexplained or challenging. Typically, contours crossing large regions of uniform intensity distribution would be identified as unexplained by our model. It is important to note that having an unexplained segmentation does not imply that this segmentation is wrong. It simply means that other priors besides those of smoothness and intensity are required to understand its boundaries.

4.2.4 Use case

We believe our approach is particularly interesting when dealing with a whole set of segmentations. For instance, say we are given a set of images with corresponding annotations. The comparison of adequacies between S and M for all images allows the detection of atypical cases which behave differently from the majority of the distribution, and for which a visual inspection might be worthy. On the contrary, applying the method on a single image is not the ideal use case as the analysis of the result is difficult without any comparison with similar images.

Our approach is unsupervised, generic, and based on few simple assumptions. However this comes with intrinsic limitations. For instance, any irrelevant contour following visible intensity boundaries will not be considered as suspicious. This limitation is common to all previously proposed unsupervised methods. More generally, the proposed method is not intended to return all erroneous segmentations inside a dataset (which is expected from a supervised approach) but instead to extract some suspicious cases when limited information is available.

4.3 Method

In this section, we review the details of our probabilistic model. We consider a binary image segmentation problem for isolating a single structure from an image I made of N voxels in a grid of dimension D ($D = 2, 3$) having intensity $I_n \in \mathbb{R}^v$, $n = 1, \ldots, N$, where $v \geq 1$ ($v = 1, 3$ and 4 in practice). We introduce for each voxel a binary hidden random variable $Z_n \in \{0, 1\}$ with $Z_n = 1$ if voxel n belongs to the structure of interest.

4.3.1 Mixtures of multivariate Student's t-distributions

Appearance models of the foreground and background regions of S are defined respectively by the two image likelihoods $p(I_n|Z_n = 1, \theta_I^1)$ and $p(I_n|Z_n = 0, \theta_I^0)$ where θ_I^0, θ_I^1 are parameters governing those models. In this chapter, we consider generic parametric appearance models as variational mixtures of multivariate Student's t-distributions [Archambeau & Verleysen, 2007]. The Student's t generalizes the Gaussian distribution with heavy tails and leads to robust mean and covariance estimates. The number of components in the mixture is automatically estimated by using a sparsity-inducing Dirichlet prior over the mixture proportions which automatically prunes the components with a small number of samples. Finally, we introduce the appearance probability ratio r_n defined as:

$$r_n(I, \theta_I^0, \theta_I^1) \triangleq \frac{p(I_n|Z_n = 1, \theta_I^1)}{p(I_n|Z_n = 0, \theta_I^0) + p(I_n|Z_n = 1, \theta_I^1)}, \tag{4.1}$$

which is the posterior label probability with a non-informative prior ($p(Z_n = 1) = 0.5$).

4.3.2 Spatial smoothness prior

The spatial smoothness prior allows the spatial organization between voxels to be taken into account and a certain degree of continuity to be enforced. To this end, different strategies can be employed. In this chapter, we propose to compare one discrete prior

(MRF) with two continuous priors (GLSP and FDSP), the third one being novel to the best of our knowledge.

MRF prior

The classical MRF formulation relies on labels of neighbouring voxels. In a binary segmentation problem, a natural way to enforce spatial smoothness is the Ising model. Assuming β to be the hyperparameter of the MRF, the label prior probability is given by:

$$p(Z|\beta) = \frac{1}{T(\beta)} \exp\left\{ \frac{\beta}{2} \sum_{i=1}^{N} \sum_{j \in \delta_i} Z_i Z_j \right\}, \tag{4.2}$$

where δ_i are the neighbouring voxels of i and $T(\beta)$ is the partition function. In practice, we consider 4- and 6-connectivity neighborhoods for 2D and 3D images, respectively. The value of β represents the strength of association between neighbouring voxels: $\beta = 0$ corresponds to a model with no spatial prior, while large positive values encourage neighbouring voxels to have the same label. The Ising model may be replaced by an image contrast sensitive prior as performed for instance in the GrabCut algorithm [Rother et al., 2004].

The computation of the partition function $T(\beta)$, needed for an automatic estimation of the model's hyperparameter β, requires considering all possible configurations of the MRF which is not computationally tractable for large lattices. Therefore, β has to be fixed by the user.

Generalized Linear Spatial Prior

In [Audelan & Delingette, 2019], we proposed a continuous label prior denoted by Generalized Linear Spatial Prior (GLSP) to enforce the spatial continuity. The prior is defined through a generalized linear model of spatially smooth functions. More precisely, the prior probability $p(Z_n = 1)$ is defined as a Bernoulli distribution whose parameter is a *spatially random* function specified as a generalized linear model:

$$p(Z_n = 1|\mathbf{W}) = \sigma\left(\sum_{l=1}^{L} \Phi_l(\mathbf{x}_n) w_l \right), \tag{4.3}$$

where $\mathbf{x}_n \in \mathbb{R}^D$ is the voxel position in an image of dimension D and the link function $\sigma(u)$ is the sigmoid function $\sigma(f) = 1/(1 + \exp(-f))$.

The basis $\{\Phi_l(\mathbf{x})\}$ are L functions of space, typically radial basis functions (for instance, Gaussian functions) defined on a regular grid, and $w_l \in \mathbf{W}$ are weights considered as random variables. Thus the prior probabilities of two geometrically close voxels are related to each other through the smoothness of the function $f(\mathbf{x}_n) = \sum_{l=1}^{L} \Phi_l(\mathbf{x}_n) w_l = \Phi_n^T \mathbf{W}$, writing $\Phi_n^T = [\Phi_1(x_n), \cdots, \Phi_L(x_n)]$.

The smoothness of the label prior $\sigma\left(f(\mathbf{x}_n)\right)$ depends on the choice of the L basis functions $\{\Phi_l(\mathbf{x})\}$ which are commonly uniformly spread over the image domain. The key parameters are the spacing between the basis centers, the standard deviations (or radii) r of the Gaussian functions and the position of the origin basis. Together, they

influence the amount of smoothing brought by the label prior, large spacing and standard deviations leading to smoother prior probability maps.

To obtain a robust description, the weight vector $\mathbf{W} = [w_1, \ldots, w_L]^T$ is fitted with a zero mean Gaussian prior parameterized by the diagonal precision matrix $\alpha \mathbf{I}_L$: $p(\mathbf{W}) = \mathcal{N}(0, \alpha^{-1}\mathbf{I}_L)$. Finally, a non-informative prior is chosen for α, $p(\alpha) \propto 1$. In contrast to the MRF formulation, a Bayesian inference of the hyperparameter is possible here, as shown in section 4.4.2.

Finite Difference Spatial Prior

As a third regularization strategy, we introduce in this chapter the Finite Difference Spatial Prior (FDSP). The prior probability $p(Z_n = 1)$ is again defined as a Bernoulli distribution whose parameter belongs to a spatially smooth random field:

$$p(Z_n = 1|\mathbf{W}) = \sigma\left(w_n\right) , \tag{4.4}$$

where $\sigma(u)$ is once more the sigmoid function. The smoothness of the label field is caused by a prior applied to the vector $\mathbf{W} = [w_1, \ldots, w_n]^T$ penalizing the squared norm of its derivatives of order p:

$$p(\mathbf{W}|\alpha) = \frac{1}{T(\alpha)} \exp\left(-\alpha \sum_{n=1}^{N} \left\|\Delta_p(w_n)\right\|^2\right) , \tag{4.5}$$

where $\Delta_p(w_n)$ is the p order central finite difference operator at w_n and $T(\alpha)$ is the normalization factor. The quantity $\Delta_p(w_n)$ is a tensor of order p approximating the p-order derivatives of the scalar field defined by w_n. Since the function $h(x) = \left\|\Delta_p(x)\right\|^2$ is 2-homogenous, we know that the normalization factor has the form $T(\alpha) = c\alpha^{-N/2}$ where c is constant independent of α [Pereyra et al., 2015]. One can easily show that $p(\mathbf{W}|\alpha)$ is a zero mean Gaussian distribution whose precision matrix consists of difference operators. The value of the parameter α controls the amount of the spatial regularization applied to the weights $\mathbf{W}$.

In this chapter, we consider only first order derivatives ($p = 1$) corresponding to the discretization of the Dirichlet energy. In that case Eq. 4.5 is written:

$$p(\mathbf{W}|\alpha) = c\alpha^{\frac{N}{2}} \exp\left(-\frac{\alpha}{4} \sum_{n=1}^{N} \sum_{d=1}^{D} (w_{\delta_d(n+1)} - w_{\delta_d(n-1)})^2\right) , \tag{4.6}$$

where $\delta_d(n + i)$ represents the neighbor of index i of voxel n in the dimension d.

The graphical models of the different segmentation frameworks are shown in Fig. 4.2.

4.3.3 Implementation

The algorithm scheme is similar regardless of the spatial prior (Fig. 4.3). The first remark is that it is of little interest to work with the whole image and computationally inefficient. The analysis is thus restricted inside a narrow band of width typically between 10 and 30 voxels defined around the boundaries of the foreground region of the input segmentation S.

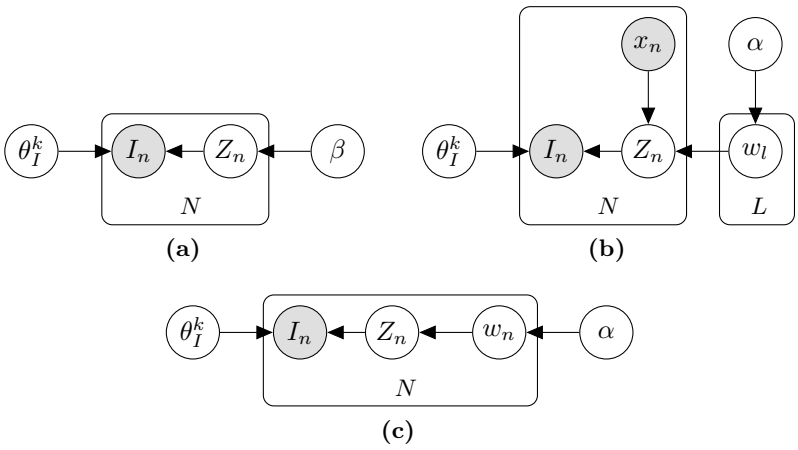


Fig. 4.2: Graphical model of the framework with a discrete MRF prior (4.2a), a GLSP prior (4.2b) or a FDSP prior (4.2c).

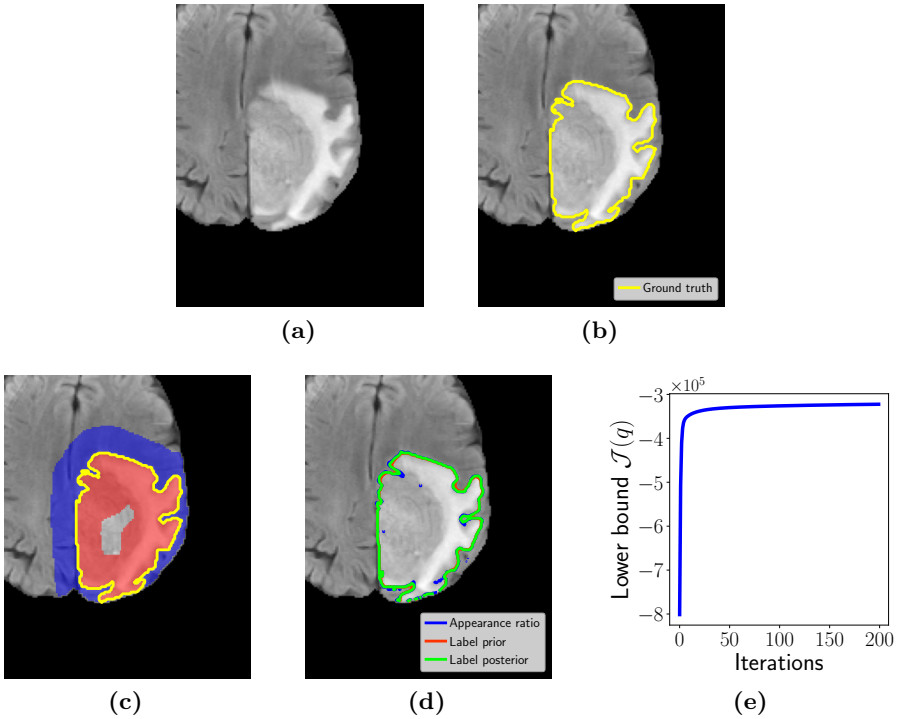


Fig. 4.3: Quality control workflow on a glioblastoma segmentation from the BRATS 2017 dataset with FDSP regularization. (4.3a) Original image. (4.3b) Input segmentation S. (4.3c) Narrow band along the ground truth boundary with foreground (in red) and background (in blue) regions. (4.3d) Appearance ratio, label spatial prior and label posterior. (4.3e) Evolution of the lower bound $\mathcal{J}(q)$.

The method starts with the estimation of the appearance probability ratio r_n for each voxel n. Two variational mixture models of Student's t-distributions are fitted, one for the foreground region of S and the other for the background.

Once the r_n are known, the second step is to compute the posterior $P(Z_n|I_n)$ which involves solving an inference problem depending on the choice of spatial prior. After convergence of the probabilistic model, a new segmentation M is generated by thresholding the posterior $p(Z_n|I_n)$ at the level 0.5.

4.4 Probabilistic inference

4.4.1 MRF regularization

A classical way to maximize the log likelihood $\log p(I)$ with an MRF prior is to use variational inference with a mean field approximation [Ambroise & Govaert, 1998; Roche et al., 2011]. The label posterior distribution $q(Z)$ is assumed to factorize as $\prod_i q_i(Z_i)$, which leads to the following fixed-point equation for voxel i at iteration $m+1$:

$$q_{ip}^{m+1} = \frac{r_i \exp\{\beta \sum_{j\in\delta_i} q_{jp}^m\}}{\sum_{k=0}^{1} r_i^k (1-r_i)^{1-k} \exp\{\beta \sum_{j\in\delta_i} q_{jk}^m\}}\,, \tag{4.7}$$

where $p \in \{0,1\}$, q_{ik} represents $q_i(Z_i = k)$, r_i is the appearance probability ratio for voxel i and β is fixed by the user.

4.4.2 GLSP regularization

A type-II maximum likelihood approach is used to estimate the model parameters. A Gaussian approximation for the weights posterior distribution is found by computing a Laplace approximation through iterative reweighted least squares. The parameter α is then updated by maximizing the marginal likelihood. We refer to the original paper for more details [Audelan & Delingette, 2019].

4.4.3 FDSP regularization

We propose a variational inference scheme to estimate prior and hyperprior parameters $U = \{Z, \mathbf{W}, \alpha\}$. Variational inference approximates the true posterior $p(U|I)$ by a chosen family of distributions $q(U)$. Maximizing the data log likelihood $\log p(I)$ implies minimizing the Kullbach-Leibler divergence between $q(U)$ and $p(U|I)$ or equivalently maximizing the lower bound $\mathcal{L}(q)$:

$$\log p(I) = \underbrace{\int_U q(U) \log \frac{p(I,U)}{q(U)} dU}_{\mathcal{L}(q)} + \mathrm{KL}\left[q(U)||p(U|I)\right]\,. \tag{4.8}$$

We assume that the approximation of the posterior can be factorized as $q(U) = q_Z(Z)q_{\mathbf{W}}(\mathbf{W})q_\alpha(\alpha)$. The lower bound can thus be re-written as:

$$\begin{aligned}\log p(I) \geqslant \mathcal{L}(q) = &\sum_Z \int_\alpha \int_{\mathbf{W}} q_Z(Z)q_{\mathbf{W}}(\mathbf{W})q_\alpha(\alpha) \\ &\log \frac{p(I|Z)p(Z|\mathbf{W})p(\mathbf{W}|\alpha)}{q_Z(Z)q_{\mathbf{W}}(\mathbf{W})q_\alpha(\alpha)} d\mathbf{W}d\alpha\,.\end{aligned} \tag{4.9}$$

We can further expand the factors defining the joint probability: $p(I|Z) = \prod_n r_n^{Z_n}(1-r_n)^{1-Z_n}$. The spatial prior $p(Z_n|\mathbf{W})$ can be likewise written as $[\sigma(w_n)]^{Z_n}[\sigma(-w_n)]^{1-Z_n}$ and the weights prior $p(\mathbf{W}|\alpha)$ is given by (4.6) for first order derivatives.

However, the right hand side of (4.9) is intractable because the spatial prior does not belong to the exponential family (due to the sigmoid function). As an alternative to the Laplace approximation, we use a local variational bound as introduced in [Jaakkola & Jordan, 2000] in the context of logistic regression. In this case, we replace the sigmoid function with a well-chosen lower bound: $\sigma(x) \geqslant g(x,\xi) = \sigma(\xi)\exp\left[(x-\xi)/2 - \lambda(\xi)(x^2-\xi^2)\right]$. ξ is a variational parameter and $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$. The spatial prior $p(Z|\mathbf{W})$ can thus be approximated by $F(Z,\mathbf{W},\xi) = \prod_n \left[g(w_n,\xi_n)\right]^{Z_n}\left[g(-w_n,\xi_n)\right]^{1-Z_n}$. This approximation leads to a new lower bound $\mathcal{J}(q)$ on the lower bound $\mathcal{L}(q)$:

$$\begin{aligned}\log p(I) \geqslant \mathcal{L}(q) \geqslant \mathcal{J}(q) = &\sum_Z \int_\alpha \int_{\mathbf{W}} q_Z(Z)q_{\mathbf{W}}(\mathbf{W})q_\alpha(\alpha) \\ &\log \frac{p(I|Z)F(Z,\mathbf{W},\xi)p(\mathbf{W}|\alpha)}{q_Z(Z)q_{\mathbf{W}}(\mathbf{W})q_\alpha(\alpha)} d\mathbf{W}d\alpha\,.\end{aligned} \tag{4.10}$$

This new lower bound $\mathcal{J}(q)$ is now tractable and the optima q^* for each of the variational posteriors can be derived by variational calculus (See appendix B.2 for details of the derivations). $q_Z^*(Z)$ is therefore given by $q_Z^*(Z) = \prod_n \eta_{n1}^{Z_n}\eta_{n0}^{1-Z_n}$ with $\eta_{nk} = \rho_{nk}/\sum_k \rho_{nk}$ for $k \in \{0,1\}$ and:

$$\begin{aligned}\rho_{nk} = r_n^k(1-r_n)^{1-k}\sigma(\xi_n)\exp\Bigg[&(-1)^{1-k}\frac{\mathbb{E}[w_n]}{2} \\ &-\frac{\xi_n}{2} - \lambda(\xi_n)(\mathbb{E}[w_n^2] - \xi_n^2)\Bigg]\,.\end{aligned} \tag{4.11}$$

By further assuming that $q_{\mathbf{W}}(\mathbf{W}) = \prod_n q_{w_n}(w_n)$, the variational optimization for $q_{w_n}(w_n)$ yields a normal distribution of the form $q_{w_n}^*(w_n) = \mathcal{N}(\mu_{w_n}, \mathbf{\Sigma}_{w_n})$. A fixed-point equation is found for updating the mean. For first order derivatives, we have:

$$\mathbf{\Sigma}_{w_n} = \left[2\lambda(\xi_n) + 2\sum_d \frac{\alpha}{2}\right]^{-1}, \tag{4.12}$$

$$\mu_{w_n} = \mathbf{\Sigma}_{w_n}\left[\eta_{n1} - \frac{1}{2} + \frac{\alpha}{2}\sum_d \left(\mu_{w_{\delta_d(n+2)}} + \mu_{w_{\delta_d(n-2)}}\right)\right]. \tag{4.13}$$

The variational posterior $q_\alpha(\alpha)$ is assumed to be a Dirac distribution which leads to the following update:

$$\alpha^{-1} = \frac{1}{2N} \sum_n \sum_{d=1}^{D} \mathbb{E}\left[\left(w_{\delta_d(n+1)} - w_{\delta_d(n-1)} \right)^2 \right] . \tag{4.14}$$

Finally, following [Bishop, 2006], maximizing (4.10) with respect to ξ_n gives an update formula of the form:

$$\xi_n^2 = \mathbb{E}[w_n^2] . \tag{4.15}$$

To compute (4.11), (4.14) and (4.15), we need the expectations $\mathbb{E}[w_n]$, $\mathbb{E}\left[\left(w_{\delta_d(n+1)} - w_{\delta_d(n-1)} \right)^2 \right]$ and $\mathbb{E}[w_n^2]$ with respect to the variational distribution q_{w_n}. They can be easily evaluated to give $\mathbb{E}[w_n] = \mu_{w_n}$, $\mathbb{E}[w_n^2] = \mathbf{\Sigma}_{w_n} + \mu_{w_n}^2$ and $\mathbb{E}\left[\left(w_{\delta_d(n+1)} - w_{\delta_d(n-1)} \right)^2 \right] = \mu_{w_{\delta_d(n+1)}}^2 + \mu_{w_{\delta_d(n-1)}}^2 - 2\mu_{w_{\delta_d(n+1)}} \mu_{w_{\delta_d(n-1)}} + \mathbf{\Sigma}_{w_{\delta_d(n+1)}} + \mathbf{\Sigma}_{w_{\delta_d(n-1)}}$.

After convergence, the variational distribution $q_Z(Z)$ gives an approximation to the posterior label probability $p(Z_n = 1 | I, \mathbf{W})$, which combines prior and intensity likelihoods. Finally, the maximum a posteriori estimate of the segmented structure is obtained as the isosurface $p(Z_n = 1 | I, \mathbf{W}) = 0.5$.

This approach has some advantages in comparison with the first two. First, it allows an automatic estimation of all its parameters. For the MRF, the user needs to fix β and for the GLSP, the layout of the basis functions and their radii are also user-defined. Moreover, a lower bound (Fig. 4.3e) on the marginal likelihood can be computed in this case, which can be used to monitor the convergence and is helpful to compare segmentation results. The computation of the lower bound is given in appendix B.3.

4.5 Results

4.5.1 Datasets

The proposed method was evaluated on four publicly available datasets: the BRATS 2017 training and validation datasets [Menze et al., 2015], the LIDC dataset [Armato III et al., 2011], the training data from the MSSEG challenge [Commowick et al., 2018] and finally the COCO 2017 validation dataset [Lin et al., 2014].

The BRATS 2017 datasets consist of multisequence preoperative MR images of patients diagnosed with malignant brain tumors. It includes 285 patients for the training dataset and 46 for the validation set. Four MR sequences are available for each patient: T1-weighted, post-contrast (gadolinium) T1-weighted, T2-weighted and FLAIR. All the images have been pre-processed: skull-stripped, registered to the same anatomical template and re-sampled to 1 mm^3 resolution. Ground truth segmentations of the brain tumors are provided only for the training set.

The LIDC dataset comprises 1018 pulmonary CT scans with 0.6 mm to 5.0 mm slice thickness. The in-plane pixel size ranges from 0.461 mm to 0.977 mm. Each scan was reviewed by 4 radiologists who annotated lesions of sizes ranging from 3 mm to 30 mm. Annotations include localization and manual delineations of the nodules. Up to 4 segmentations can be available for the same nodule, depending on the number of

radiologists who considered the lesion to be a nodule. In this chapter, all scans were first re-sampled to 1 mm^3 resolution as pre-processing step, and we restrict the analysis to nodules of diameter above 20 mm, i.e. 309 segmentations.

The MSSEG training dataset contains MR data from 15 multiple sclerosis (MS) patients. Manual delineations of lesions were performed on the FLAIR sequence by seven experts.

Finally, COCO is a large-scale object detection and segmentation dataset of real world images. The 2017 validation set contains 5000 images with 80 object categories, ground truth object classification, object localization and segmentation. To annotate such a large number of images, the authors resorted to a crowd-sourcing annotation pipeline.

4.5.2 Unsupervised indices

As discussed in section 4.1, different indices have been proposed in prior works for unsupervised segmentation evaluation. We selected 4 of them in order to provide a qualitative and quantitative comparison with our approach. They all involve the computation of 2 metrics, the former measuring the intra-region uniformity while the latter gives an estimate of the inter-region disparity.

Three out of the four indices are taken from [Zhang et al., 2008]: Zeb, η and F_{RC}. The last one was introduced in [Johnson & Xie, 2011] and is denoted by GS in this chapter. Formula are given in appendix B.1.

4.5.3 Setting hyperparameters

Width of the narrow band

As noted in section 4.3.3, the analysis is restricted to a narrow band alongside the input segmentation's contour. The width of this narrow band controls the extent of the region taken into account for learning the appearance models for both background and foreground and fitting the regularization model.

We assessed the sensitivity of the results to this hyperparameter on the BRATS training set and LIDC dataset. We applied the algorithm for several narrow band widths using FDSP as a spatial prior. Different ASE values were obtained for each segmentation depending on the narrow band setting. We then analysed the stability of the sets made by the 40 segmentations with the largest ASE values by computing pairwise intersection over union (IoU) coefficients. A value of 1.0 indicates that the 40 images are the same for a pair of narrow band widths. The outcome is shown in Fig. 4.4.

While the sets from the BRATS training set are rather stable, those from the LIDC dataset show some variability. An explanation of the sensitivity of LIDC segmentations to the narrow band width can be found in Fig. 4.5. If the narrow band is too wide, the high intensity differences between the pleura and the lung parenchyma lead appearance models of nodules close to the pleura to leak outside the lung.

In brief, the sensitivity of the algorithm with respect to the narrow band's width varies from case to case. As the computation time is not a bottleneck for FDSP regularization, we propose in practice to perform the analysis with different width settings and then choose the one leading to the most stable and reasonable results.

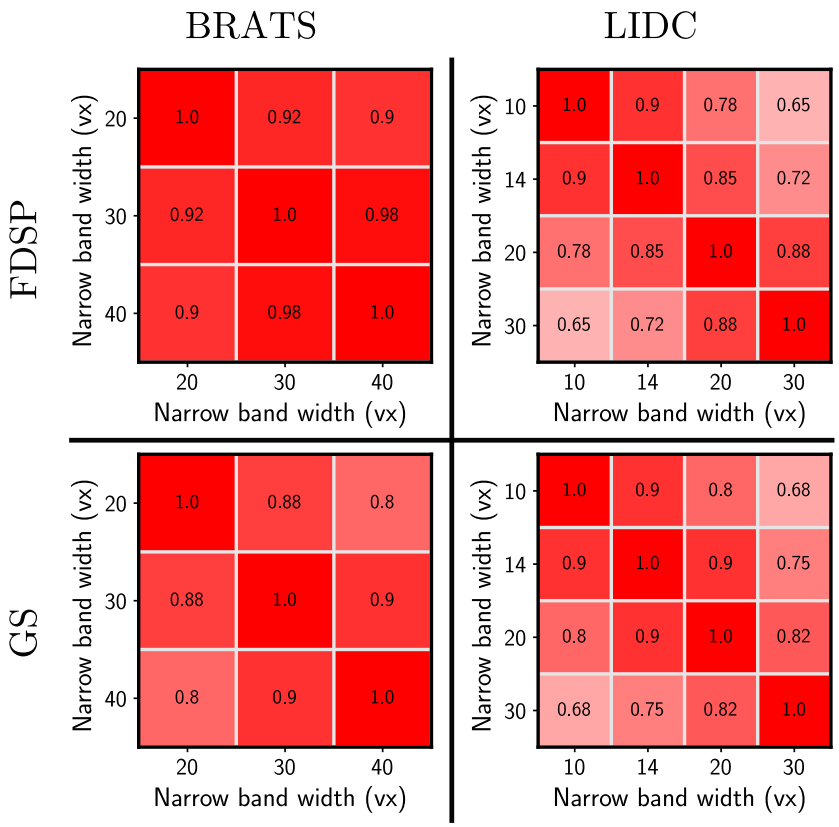


Fig. 4.4: Sensitivity analysis of the narrow band width for the BRATS and LIDC datasets with our approach using FDSP as a spatial prior (first row) or with the unsupervised indicator *GS* (second row). Matrices show IoU scores computed between the sets made of the 40 segmentations with largest ASE (first row) or largest *GS* score (second row).

We would also like to underline that previously published unsupervised indices are likewise sensitive to the width of the narrow band. An example is shown in Fig. 4.4 for the indicator *GS*. In order to provide a fair comparison between approaches, the computations of the selected unsupervised indices are always performed on the same narrow band as the one used for our method.

Other hyperparameters

Among the parameters that need to be defined by the user is the number of components for the mixtures of multivariate Student's t-distributions. It is fixed to 7 in all our experiments. This parameter is not so sensitive as unnecessary components will be pruned by the Dirichlet prior and removed from the model.

The number of remaining parameters depends on the chosen spatial prior. For an MRF prior, the user needs to provide a value for β, which controls the strength of the regularization. We tested 3 values for this hyperparameter throughout our experiments: 0.2, 1 and 3. For a GLSP regularization, the user has to define a dictionary of basis functions whose key parameters are the step between each basis function and their radii. They likewise control the amount of regularization. In this chapter, we set the step to 6 vx and the radius to 17 vx, except for the LIDC dataset for which the step was set

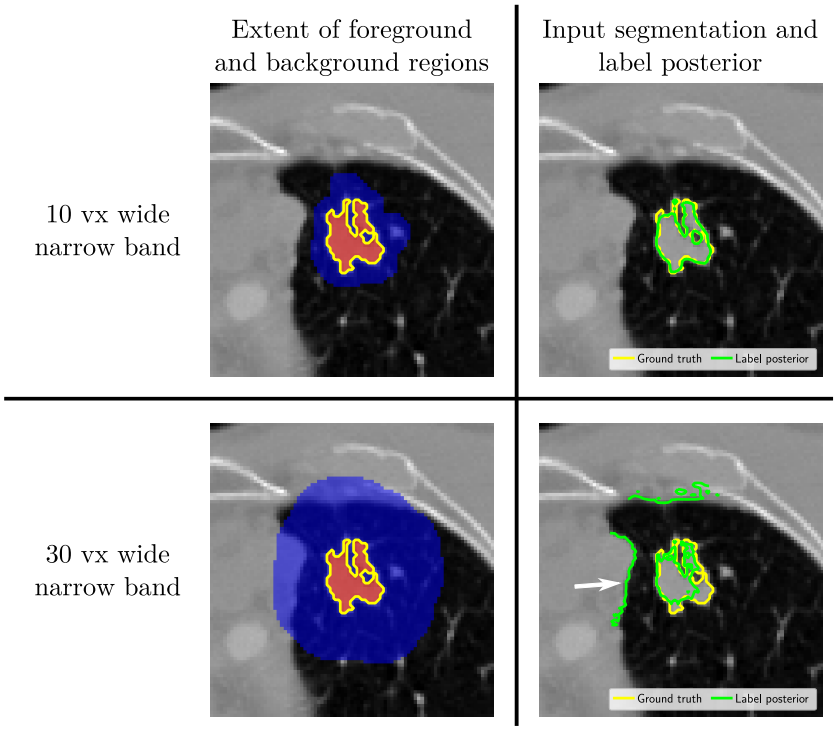


Fig. 4.5: Example of a nodule segmentation from LIDC where the result of the quality assessment is different depending on the narrow band width. If too large, the appearance model of the foreground leaks inside the pleura leading to an irrelevant result.

to 4 vx and the radius to 12 vx. Finally, for the regularization using an FDSP prior, no further parameter needs to be set by the user as the model's hyperparameters are all learnt automatically, which is a great advantage in comparison with the first two approaches.

4.5.4 Qualitative analysis

In the case of segmentations produced by human raters, possibly with the help of interactive annotation tools, it is very useful to be able to rank segmentations, highlight potentially difficult segmentations and track possible errors in large databases.

In this section we present some results from two datasets of medical images, whole brain tumor segmentations from the BRATS 2017 training set and pulmonary nodule segmentations from the LIDC dataset. On average, one minute is required to complete the quality control workflow for a 3D image from the BRATS dataset using an MRF or FDSP regularization. The inference time increases to 4 minutes for a model with a GLSP prior. The computation time of course also depends on the size of the segmented structure and on the extent of the narrow band.

Computation of the ASE for each segmentation allows the distribution for the whole dataset to be drawn. Histograms obtained with FDSP regularization are shown in Fig. 4.6. They present a similar shape, with a short left tail, a single peak and a heavier right tail. Cases in the right tail isolated from the rest of the distribution are atypical and possibly include errors. Samples from the left and right tail are shown in Figs. 4.7 and

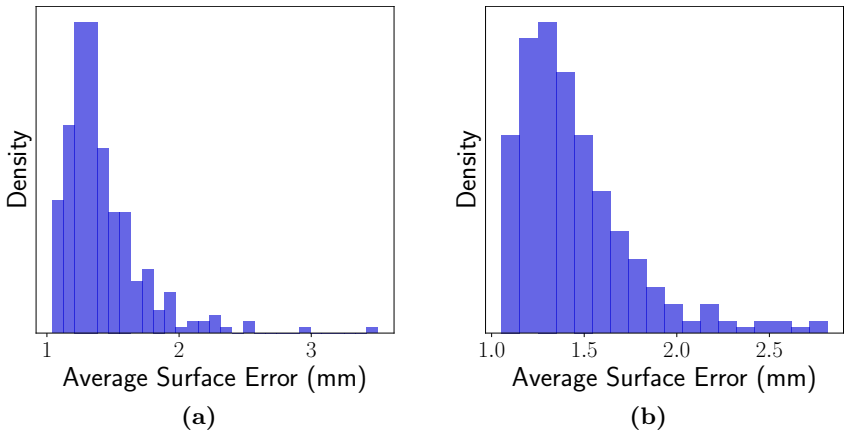


Fig. 4.6: ASE distributions for the analysis of ground truth segmentations from the BRATS (4.6a) and LIDC datasets (4.6b). Samples from the left tail are identified as explained by the model while samples from the right tail are classified as challenging.

4.8, respectively. For both datasets, cases with larger ASE are clearly more challenging than the cases taken from the left tail.

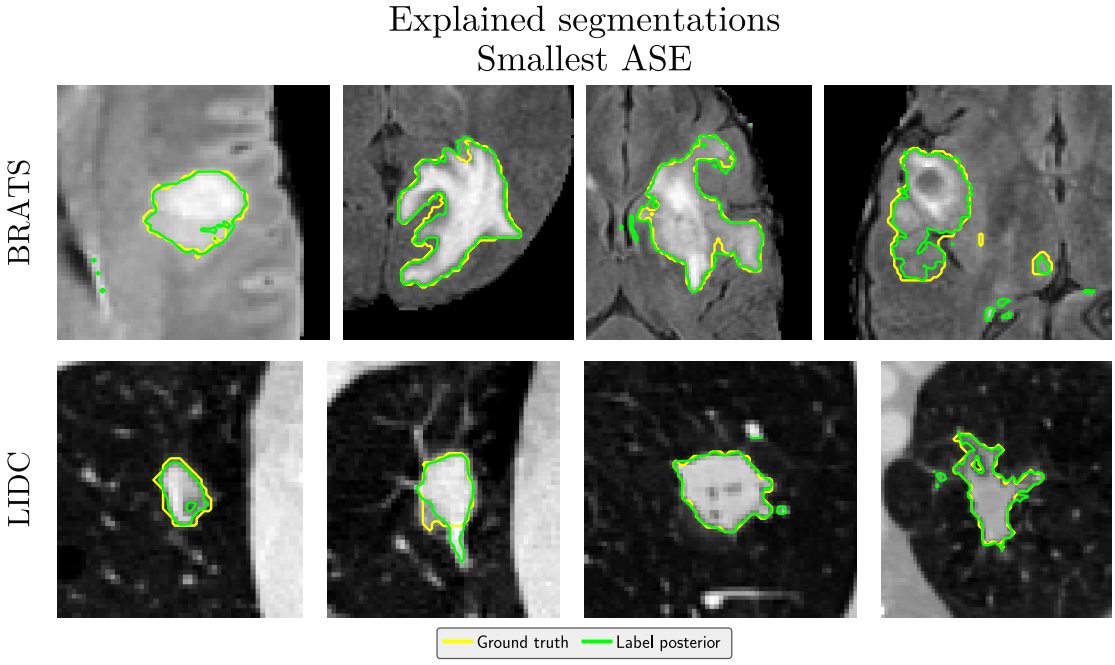


Fig. 4.7: Segmentations with the smallest ASE taken from the left tail of the distributions. Cases are ranked according to their ASE value (Largest values to the right) and slices with largest ground truth area are shown. The width of the narrow band is 30 vx for BRATS and 10 vx for LIDC.

Furthermore, one can see that contours in the right tail samples from BRATS are more irregular and that intensity variations in some regions are very weak making their accuracy questionable. Those contours were probably extracted through thresholding instead of being manually drawn as was permitted in the annotation process [Jakab, 2012;

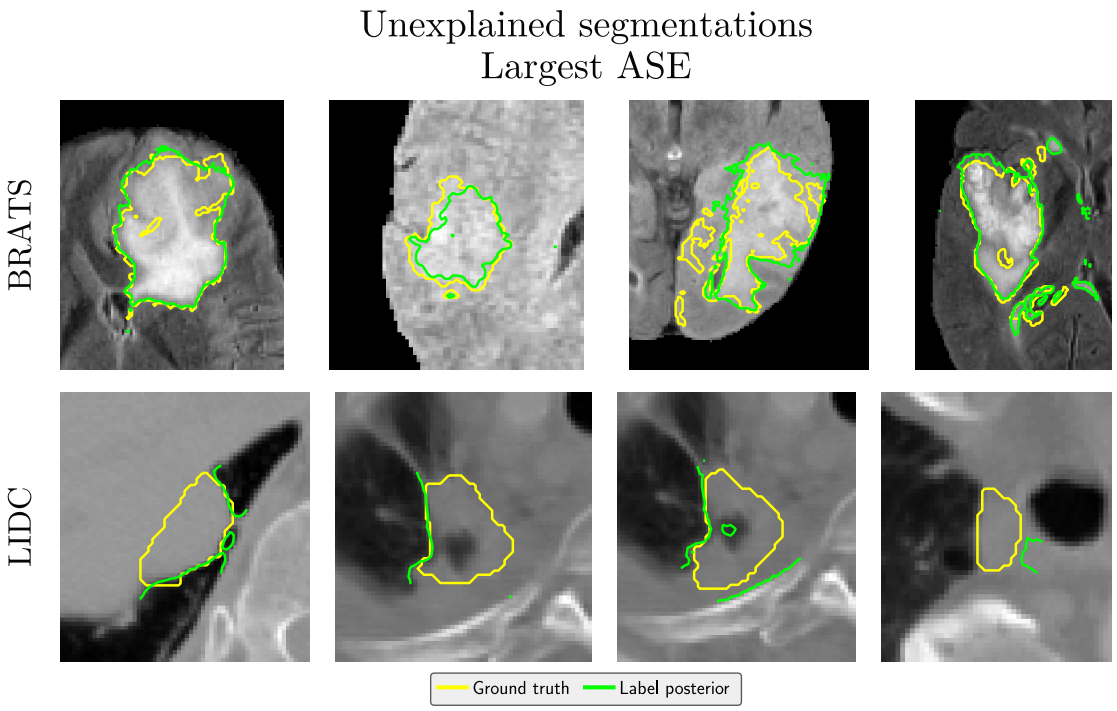


Fig. 4.8: Segmentations with the largest ASE taken from the right tail of the distributions. Cases are ranked according to their ASE value (Largest values to the right) and slices with largest ground truth area are shown. The width of the narrow band is 30 vx for BRATS and 10 vx for LIDC.

Menze et al., 2015]. Similarly, some contours in the right tail samples from LIDC cross regions of uniform intensity and therefore require other priors like shape to be explained. Yet, the contours are far from obvious in some areas in comparison with the left tail samples. Therefore, our approach fulfills its role of extracting challenging, possibly suspicious, cases within a dataset.

We present in Fig. 4.9 a qualitative comparison between the spatial priors proposed for our approach and the unsupervised indices presented in section 4.5.2. Each dataset was sorted according to those indices and the 40 segmentations with largest ASE/score were extracted. The variability of this set of suspicious segmentations across unsupervised methods was studied by computing the pairwise IoU.

First, we note that our approaches and the unsupervised indices yield different sets of suspicious segmentations, as the IoU score is always less than 0.4. Furthermore, the unsupervised indices lead to inconsistent results on both datasets which make their performance highly unreliable on medical images. One possible explanation is that those methods were designed for 2D color images with large contrast and may not scale well to 3D medical images.

If we now compare the different regularization strategies proposed for our approach, we observe that the level of regularization has some impact. Indeed, there is a significant variability of results with the value of β for the MRF prior. This observation supports using the last regularization strategy proposed, the FDSP prior, as in this case all hyperparameters are learnt in a data-driven way.

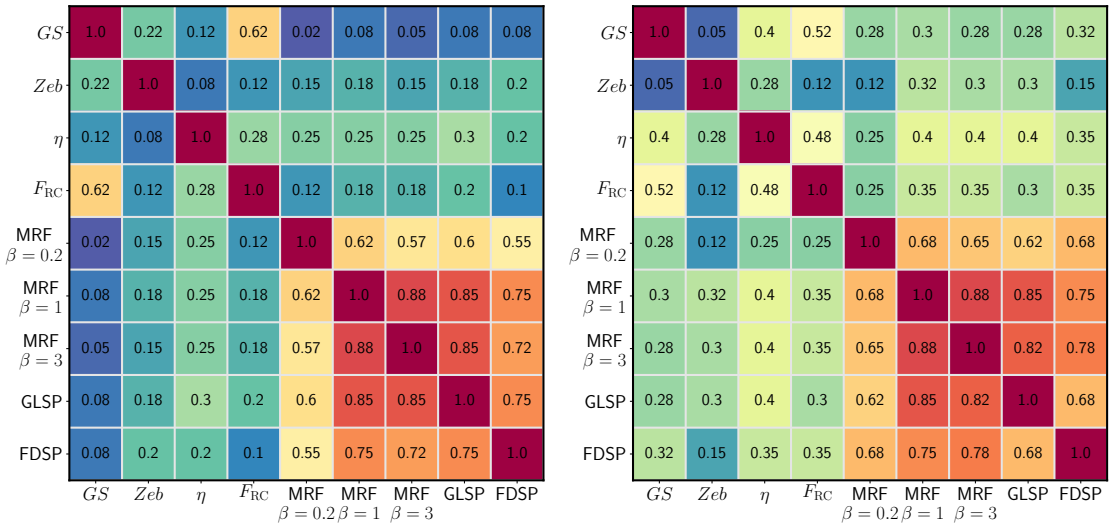


Fig. 4.9: Comparison of different approaches on the BRATS training set (4.9a) and LIDC (4.9b). The 40 segmentations with largest ASE or indicator score are compared using IoU. The width of the narrow band is 30 vx for BRATS and 10 vx for LIDC.

4.5.5 Quantitative analysis

In order to perform a quantitative comparison of the different methods, we need to have a grading of the quality of all segmentations. As they are easier to obtain for real world images than medical images, we propose to conduct this quantitative assessment on the COCO dataset which contains real world pictures with a large variability among them, with grayscale or color images, segmented structures of variable sizes and large ranges of noise level.

Quality grading process

Seven object categories from the COCO dataset were selected for the quantitative assessment: airplane, bear, dog, snowboard, couch, bed and handbag (see Fig 4.10). Each segmentation was ranked according to the different methods, leading to 9 distributions for each object category: FDSP, GLSP, MRF with 3 values of β and the 4 unsupervised indices. The width of the narrow band was set to 30 px (pixels) for all approaches. Since grading the entire set of images would have been too time-consuming, we chose to focus on the right tails of the distributions (segmentations with largest ASE or indicator score) where the suspicious cases are expected to lie.

For each distribution, we extracted segmentations from the right tail corresponding to 20% of the total distribution. If the number of extracted segmentations was larger than 40, only the 40 cases with largest ASE/score were retained. All segmentations were pooled leading to a total of 703 delineations.

Six raters were then recruited in order to grade each segmented image as good or poor through a custom application presenting the segmentations in a random order. Raters were asked to repeat the annotation twice in order to estimate the intra-rater variability. The intra-rater variability was found to be slightly lower than the inter-rater

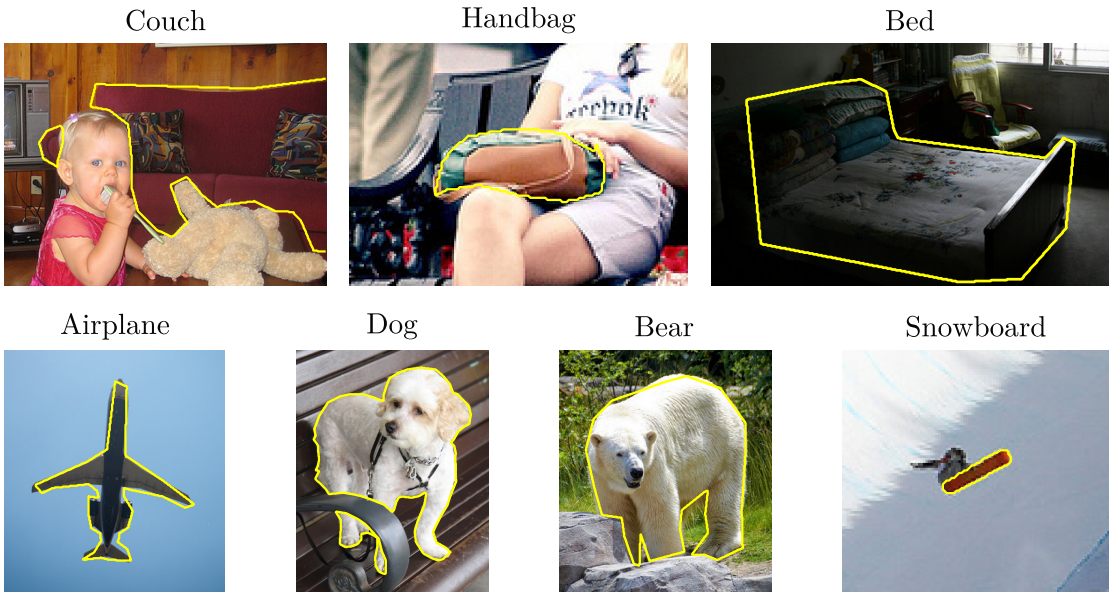


Fig. 4.10: Examples of ground truth segmentations from the COCO dataset representing the 7 selected object categories.

variability, with a mean rate of identical responses of 83% for the former and of 73% for the latter.

Performance comparison

The objective is to compare the segmentation quality among the right tails of the distributions given by the different approaches. These tails correspond to segmentations with the largest ASE or index score. The percentage of cases rated as poor by the raters strongly depends on the size of the set of segmentations extracted from the right tail of each distribution. Yet, it is useful to compare two quality control algorithms since a better algorithm is expected to have a greater proportion of segmentations annotated as poor by the raters than a worse one.

Each segmentation was assigned to a quality category, good or poor, after taking the mean across the raters' responses. Proportions of poor segmentations per approach were then derived for each object category. Distributions of these proportions over the 7 object categories are shown in Fig. 4.11. Two observations can be made. First, our approaches show competitive results as they lead to higher mean and median proportions of poor segmentations than the unsupervised indices. Second, no regularization strategy seems better than the others. In particular, variations of the value of β do not affect the results very much for the MRF.

Fig. 4.12 is obtained after pooling all object categories. To assess the robustness of the results, different thresholds are used to select the segmentations taken into account, depending on the level of agreement among raters' responses. Fisher's exact test is used to assess the difference between the proportion for a given approach and the one obtained with an FDSP prior. Three unsupervised indices η, Zeb and GS, are found to give significantly different results than our approach with FDSP regularization, regardless of the threshold. More generally, our approaches always lead to a higher percentage of poor segmentations than the indices. Again, all regularization strategies seem to be

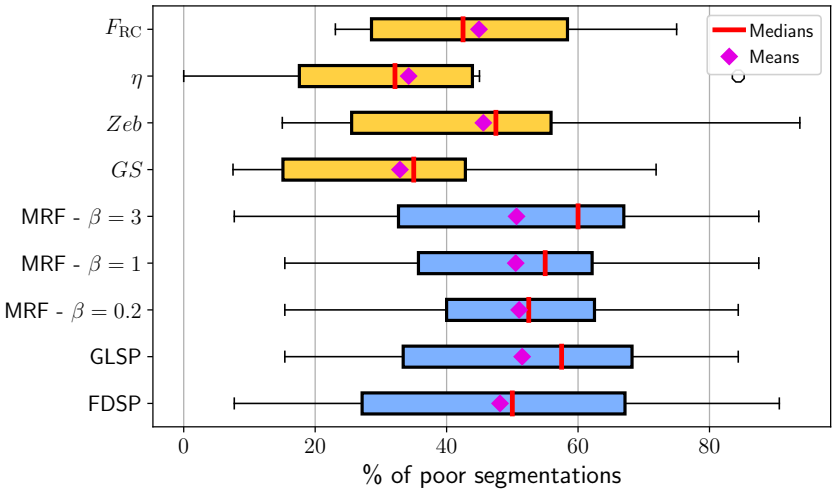


Fig. 4.11: Distribution of the proportion of poor segmentations over the 7 object categories for each approach. The mean over the raters is taken as the final label for each segmentation.

appropriate. The results seem to be stable with respect to the level of regularization enforced by β.

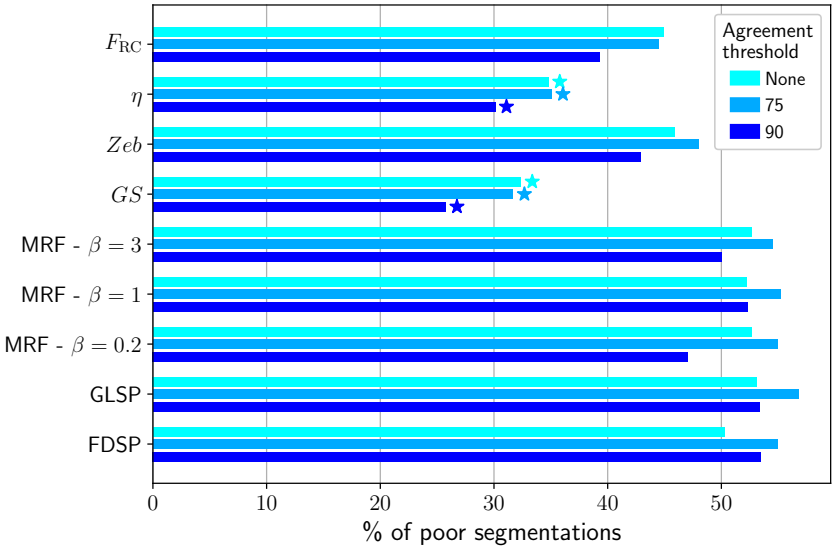


Fig. 4.12: Proportion of poor segmentations after pooling all object categories. Only segmentations with a raters' agreement above a given threshold are retained in the computation of the proportion. Results found to be significantly different from the ones given by the FDSP prior with Fisher's exact test at significance level 0.05 are marked with star symbols $\star$.

Comparison with inter-rater variability

Assessing the quality of segmentations inside a medical imaging dataset is difficult without any expert knowledge. However, some datasets provide several segmentations of the same image produced by different experts. For instance, up to four segmentations

are available for each nodule in the LIDC dataset and MS lesions in the MSSEG training dataset were delineated by seven radiologists. The inter-rater variability measures the level of agreement between the experts. It is reasonable to assume that images for which there is a low level of agreement between the experts are more challenging than others. Therefore we study in this section the relationship between inter-rater variability and the score produced by our unsupervised model.

Tab. 4.1: Values of the correlation coefficient between the inter-rater variability and the average score given by different methods on the LIDC dataset. The width of the narrow band is 10 vx and FDSP was used as a spatial prior for our model.

	Inter-rater variability		
	Avg Dice score	Avg HD	Avg 95% HD
Avg F_{RC}	0.11	0.05	-0.03
Avg Zeb	-0.34	0.12	0.2
Avg η	0.13	-0.18	-0.12
Avg GS	0.01	0.03	0.05
Avg Dice score between S and M	**0.47**	**-0.32**	**-0.39**
Avg ASE between S and M	0	0.05	0.03

Tabs. 4.1 and 4.2 show the correlation coefficient between the inter-rater variability and the Dice score or ASE produced by our model on the LIDC and MSSEG datasets, respectively. We also compare with the four unsupervised indices selected earlier. The inter-rater variability was quantified in three manners: by computing the average Dice score between all pairs of experts, the average pairwise Hausdorff distance (HD) and the average 95% percentile of the pairwise Hausdorff distance (95% HD). It is compared to the average score computed on the different raters' segmentations for each unsupervised method. For the LIDC dataset, we discarded all nodules annotated by a single radiologist, leaving a total of 87 nodules.

Tab. 4.2: Values of the correlation coefficient between the inter-rater variability and the average score given by different methods on the MSSEG dataset. The width of the narrow band is 20 vx and FDSP was used as a spatial prior for our model.

	Inter-rater variability		
	Avg Dice score	Avg HD	Avg 95% HD
Avg F_{RC}	0.17	-0.21	-0.36
Avg Zeb	-0.72	0.14	0.44
Avg η	-0.55	0.03	0.32
Avg GS	0.06	-0.19	-0.25
Avg Dice score between S and M	**0.81**	**-0.49**	**-0.7**
Avg ASE between S and M	-0.64	0.47	0.67

Better correlations are achieved on the MSSEG dataset than on the LIDC dataset. Furthermore, our approach differs significantly from the other unsupervised indices with much larger correlation values. The others (except Zeb) exhibit indeed coefficients close to zero.

We further analyse the link with inter-rater variability by showing some examples from both datasets on Fig. 4.13. The first row presents results on the MSSEG dataset, where the correlation is quite good (0.81). Case A has a high inter-rater variability and is labelled as challenging by our model (low average Dice between the inputs and the model). Indeed, only three raters out of seven considered that some lesions were visible on the slice presented in Fig. 4.13b. Moreover, the low intensity contrast does not help to understand the segmentations. On the other hand, case B is better explained by the model with a good agreement between the experts, as shown in Fig. 4.13c.

The bottom row shows poorer results on the LIDC dataset. Two contradictory cases are highlighted. The first one, case C, has a low inter-rater variability but is predicted as challenging by our model (low average Dice score between S and M). The two radiologists are indeed giving close contours (Fig. 4.13e) but it is also clear that the case is challenging according to the assumptions of our model. In the image regions highlighted by the arrows, the contours are indeed crossing areas of uniform intensity distribution, which make them more difficult to understand. On the other hand, case D is a typical case illustrating the limitations of our model (Fig. 4.13f). Raters disagree about the extent of the nodule, but all segmentations correspond to visible boundaries and match the assumptions of our model. One possible explanation for the poorer correlation obtained on the LIDC dataset is that the annotations were made in two stages, the second stage allowing radiologists to see the annotations made by the other experts in the first stage. This may have led to a decrease in inter-rater variability.

This analysis shows that in some cases, the inter-rater variability may not be a good surrogate of the difficulty of a segmentation. Raters may provide similar segmentations despite the fact that they are not close to visible boundaries (Case C in Fig. 4.13e) in the image. In that case, a low inter-rater variability is associated with a difficult segmentation.

4.5.6 Results interpretability

The previous section demonstrated how well our approach performed in extracting suspicious segmentations from a dataset in an unsupervised manner. However, it also differs from approaches proposed in the literature regarding the output of the algorithm. For instance, the unsupervised indices output only a scalar score as a ratio of 2 metrics measuring the intra-region homogeneity and the inter-region dissimilarity. In our case, the output of the algorithm is a new segmentation used as a comparison tool. Although this segmentation must not be seen as a surrogate ground truth, it can help to visually understand why a segmentation is considered atypical, that is, has a large ASE, which is not possible with the indices.

Voxels lying on the input segmentation border can thus be colored depending on their distance to the model segmentation contour, as shown in Fig. 4.14. When dealing with 3D medical images with a large number of slices, it is useful to be able to retrieve quickly the most problematic regions according to the model. Identifying the most suspicious slices is not possible with approaches outputting a simple score. Last but not least, the model segmentation could also be used as a guide for the correction of poor cases.

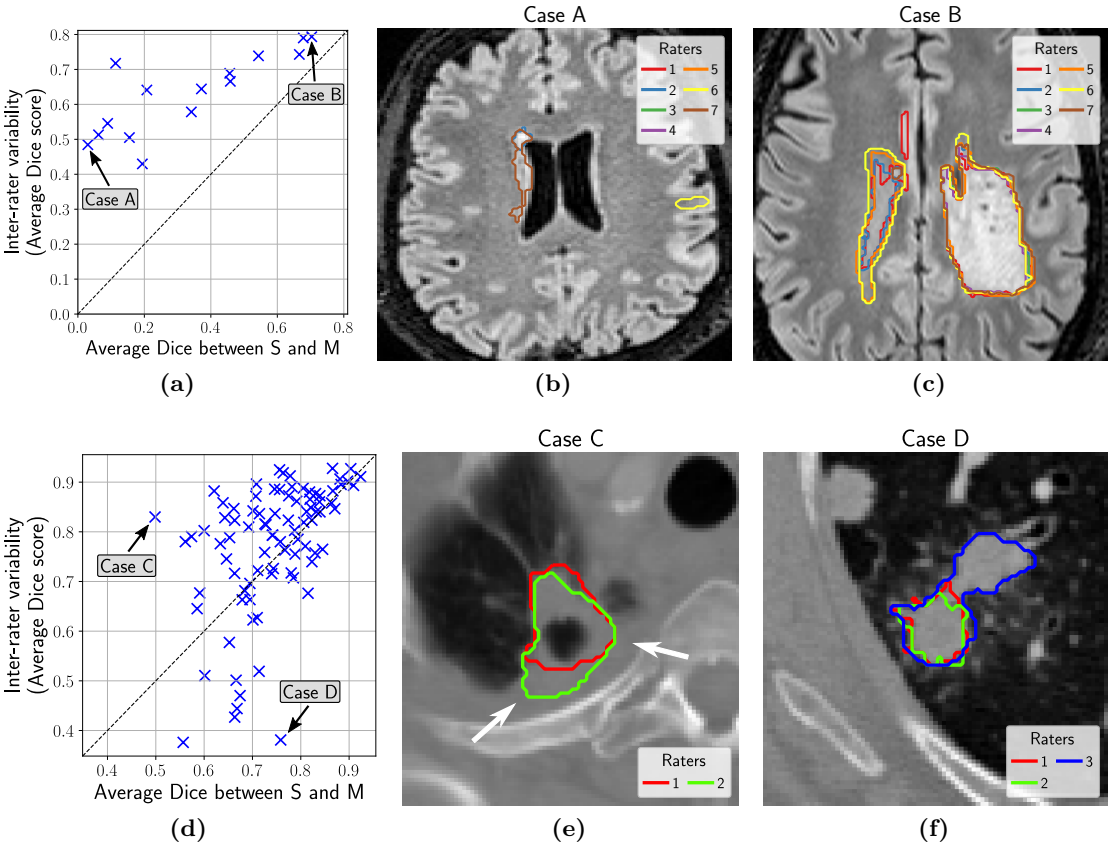


Fig. 4.13: Correlation between the inter-rater variability and the difficulty of a segmentation as predicted by our model on the MSSEG dataset (top row) and LIDC dataset (bottom row).

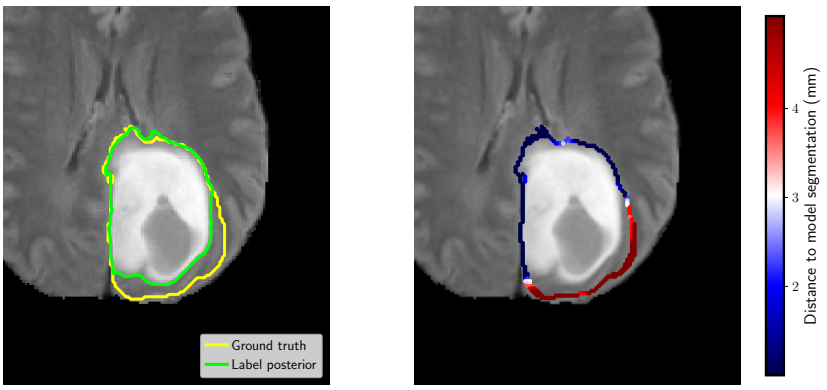


Fig. 4.14: Interpretability of the result given by our approach on a brain tumor segmentation from BRATS. (4.14a) Ground truth segmentation and label posterior given by the probabilistic model with FDSP regularization. (4.14b) Coloring of voxels lying on the ground truth border depending on their distance to the ouput of the probabilistic model.

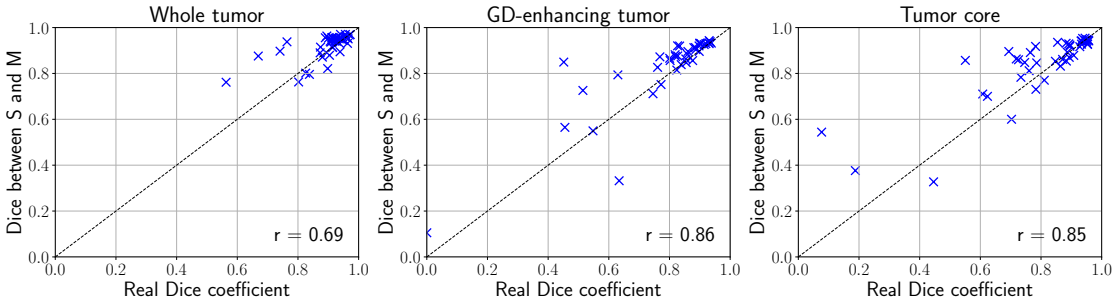


Fig. 4.15: Real Dice coefficient versus Dice score between the prediction S of the CNN and the probabilistic segmentation M with FDSP prior exhibiting good correlation. Results are shown for a narrow band width of 30 vx on 3 tumor compartments.

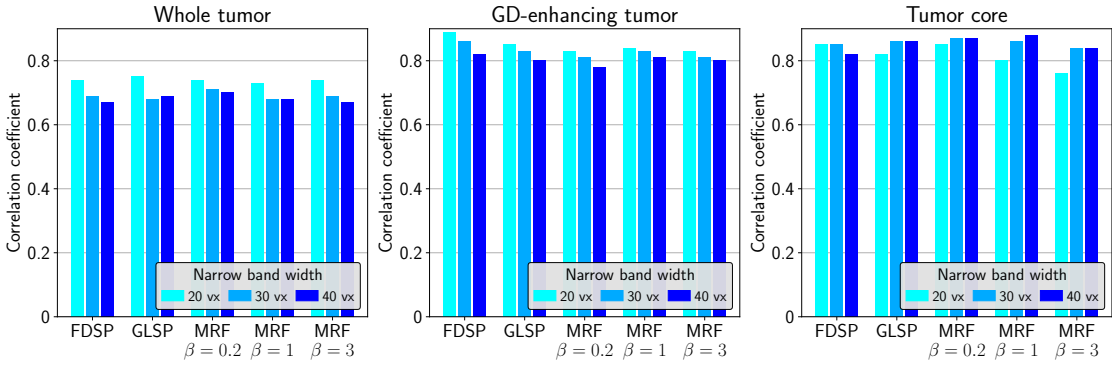


Fig. 4.16: Values of the correlation coefficient between the real Dice and estimated Dice score for different regularization strategies and different widths of the narrow band.

4.5.7 Surrogate segmentation performance

In this section we investigate if metrics estimated by our segmentation quality assessment algorithm can be correlated with the overall segmentation performance of an algorithm. In particular, we consider the segmentations generated by a convolutional neural network (CNN) detailed in [Mlynarski et al., 2019] on 46 test images of the BRATS 2017 challenge. The Dice score computed between the predicted segmentation S and the one obtained by thresholding the posterior map, M, is then compared to the true Dice index obtained by uploading the generated segmentation on the evaluation website of the challenge. In other words, we want to assess if the Dice score between S and M can be predictive of the real segmentation performance of the algorithm.

Correlations obtained with an FDSP prior on a narrow band of width 30 vx are given in Fig. 4.15 for the 3 different tumor compartments and are all above 0.69 with few outliers. Fig. 4.16 present correlation coefficients with all regularization strategies and for different values of the narrow band width. The coefficients are very similar across the approaches and are little affected by the variations of the narrow band width.

However, we do not find that this approach always predicts the performance of segmentation algorithms well. For instance, we have noticed poor predictions for most categories in the COCO dataset. This can be explained by the fact that good performance predictions can only be obtained when the segmented structure follows the model assumptions, that is, the background and foreground regions have different mixtures of Student's t-distributions.

4.5.8 Discussion

The proposed unsupervised quality control method was shown to efficiently and automatically isolate challenging or atypical segmentation cases from a whole dataset. It was shown to outperform four previously introduced segmentation quality indices on the COCO dataset. Furthermore, those four indices do not provide stable results on the LIDC and BRATS medical datasets. The proposed algorithm does not produce a classification between good or poor segmentations but rather a ranking between cases within a dataset.

The genericity of the algorithm allows it to work on any type of object category or image (2D RGB or 3D grayscale images). We demonstrated the ability of the method to handle a wide range of segmentations, from small structures (lung nodules) to large brain tumor delineations. Yet, the approach is not suited for very tiny objects since a reasonable size is required to have a reliable estimation of the intensity parameters. Also, the spatial prior is likely to wipe out the segmentation if its area is really too small. Furthermore, the genericity of the algorithm may also be considered as a limitation when focusing on a specific structure of interest. For instance, if we aim at segmenting objects from the car category on the COCO dataset, a contour perfectly following intensity boundaries but around another object category would not be identified as atypical. To this end, one would need to also monitor several specific features of that structure such as its color, size or shape, which amounts to performing a supervised quality control as in [Xu et al., 2009]. This limitation is shared by all unsupervised quality control methods.

Another limitation is the difficulty to distinguish boundaries in areas with low intensity contrast. Our method is based on mixtures of Student's t-distributions, which is already a far more general assumption than some previous unsupervised approaches that hypothesize a unique Gaussian distribution in each region [Zhang et al., 2008]. Furthermore, our Bayesian formulation integrates intensity and smoothness assumptions into a single probabilistic model, as opposed to previous unsupervised methods, which require weighting of the heterogeneity and homogeneity terms.

Different spatial regularization strategies are proposed and tested in this chapter. Quantitative assessment on COCO seems to indicate that all approaches lead to similar results. However, the FDSP prior based on derivative penalization does not require any hyperparameters to be set while keeping the computation time low, supporting its use in preference to the others.

Finally, compared to learning-based approaches such as [Kohlberger et al., 2012] or [Robinson et al., 2018] and also to previous unsupervised indices which only output a score, our method provides an explanation for the mismatches between the posterior probabilities M and the input segmentation S. This is a major advantage considering the growing importance of providing interpretable models.

4.6 Conclusion

Image segmentation is an important task in medical image analysis and computer vision. Quality control assessment of segmentations is therefore crucial, but the trend towards the generation of large databases makes any human-based monitoring onerous if not impossible. This chapter introduces a new framework for generic quality control assessment which relies on a simple and unsupervised model. It has the advantage

of not requiring a priori any knowledge about the segmented objects nor a subset of trusted images to be extracted. This is especially suited to the monitoring of manually created segmentations, where potential errors can be found, as shown by our results. Its application to segmentations generated by algorithms is also of great interest and in some cases can be used as a surrogate for segmentation performance.

The proposed generic segmentation model produces contours of variable smoothness that are mostly aligned with visible boundaries in the image. Three regularization strategies were proposed in this chapter and produced similar results. However, the prior based on derivative penalization has the great advantage of allowing an automated estimation of all hyperparameters with variational Bayesian inference, which is not possible within the classical MRF framework.

Extensive testing has been performed on different datasets containing various types of images and segmented structures, showing the ability of the method to isolate atypical cases and therefore to perform quality control assessment. Comparison with unsupervised indices from the literature proved our approach to be effective and competitive. Coping with multiple foreground labels may be an interesting extension to process multiple regions of interest jointly rather than sequentially. Finally, an interactive use of the proposed algorithm during the manual delineation of structures in images is an exciting perspective to help reduce the inter-rater variability in the context of crowdsourcing.

Acknowledgments

This work was partially funded by the French government, through the UCA$^{\text{JEDI}}$ and 3IA Côte d'Azur "Investments in the Future" projects managed by the National Research Agency (ANR) with the reference numbers ANR-15-IDEX-01 and ANR-19-P3IA-0002 and supported by the Inria Sophia Antipolis - Méditerranée "NEF" computation cluster.

Chapter 5

Robust Bayesian fusion of continuous segmentation maps

Contents

5.1 Introduction . . . 74
5.2 Robust estimate of consensus probability maps . . . 77
5.2.1 Baseline probabilistic framework . . . 77
5.2.2 Heavy-tailed distributions and scale mixture representation . 79
5.2.3 Model inference . . . 81
5.3 Mixture of consensuses . . . 85
5.3.1 Probabilistic framework . . . 85
5.3.2 Model inference . . . 85
5.4 Results . . . 87
5.4.1 Material . . . 87
5.4.2 Robust probabilistic framework . . . 89
5.4.3 Mixture of consensuses . . . 98
5.5 Discussion . . . 101
5.6 Conclusion . . . 102

The fusion of probability maps is required when trying to analyse a collection of image labels or probability maps produced by several segmentation algorithms or human raters. The challenge is to weight the combination of maps correctly, in order to reflect the agreement among raters, the presence of outliers and the spatial uncertainty in the consensus. In this chapter, we address several shortcomings of prior work in continuous label fusion. We introduce a novel approach to jointly estimate a reliable consensus map and to assess the presence of outliers and the confidence in each rater. Our robust approach is based on heavy-tailed distributions allowing local estimates of raters performances. In particular, we investigate the Laplace, the Student's t and the generalized double Pareto distributions, and compare them with respect to the classical Gaussian likelihood used in prior works. We unify these distributions into a common tractable inference scheme based on variational calculus and scale mixture representations. Moreover, the introduction of bias and spatial priors leads to proper rater bias estimates and control over the smoothness of the consensus map. Finally, we propose an approach that clusters raters based on variational boosting, and thus may produce several alternative consensus maps. Our approach was successfully tested

on MR prostate delineations and on lung nodule segmentations from the LIDC-IDRI dataset.

This chapter corresponds to the following publications:

- Robust Bayesian fusion of continuous segmentation maps. *In preparation for submission to a journal.*
- [Audelan et al., 2020] **B. Audelan**, D. Hamzaoui, S. Montagne, R. Renard-Penna and H. Delingette. Robust Fusion of Probability Maps. *In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 259–268.*

5.1 Introduction

The fusion of probability maps is necessary to solve at least two important problems related to image segmentation. The first is to establish the underlying ground truth segmentation given several binary or multi-class segmentations provided by human raters or segmentation algorithms (e.g., in the framework of multi-atlas segmentation [Sabuncu et al., 2010]). This is especially important in the medical domain, where manual contour delineations are known to suffer from potentially large inter-observer variability, due to objective factors like the image quality, but also to more subjective ones, such as the observer level of expertise [Joskowicz et al., 2019]. The generated segmentation masks might have a direct impact on clinical decisions, for example in cancer radiotherapy planning where delineation discrepancies could result in significant differences regarding the definition of the target region [Petersen et al., 2007]. Moreover, in the computer vision domain, accurate consensus estimations are needed for the performance assessment of segmentation algorithms, as comparison with expert delineations is the gold standard in the absence of physical or virtual phantoms. Indeed, the data fusion method used to build the reference can significantly impact the ranking result when comparing several segmentation algorithms [Lampert et al., 2016]. Another domain requiring robust segmentation estimation is radiomics analysis. For instance, radiomics models can be used to make predictions about a tumor. These predictions are based on features extracted from the tumor region in the image, which is typically defined by the segmentation. It has been shown that variations in the delineation of the tumor volume lead to a poor reproducibility of the radiomics results [Kocak et al., 2019], thus highlighting the importance of robust consensus estimation to mitigate the adverse effects of inter-rater variability.

The second related problem is the fusion of probability maps that are outputted by several segmentation algorithms such as neural networks. For instance, in [Wang et al., 2019], a 3D segmentation is obtained from several 2D maps generated by a neural network using a statistical fusion approach. Similarly, data fusion is needed in [Tang et al., 2021] to aggregate results obtained at a patch level into a final segmentation. Finally, it has also been shown experimentally that combining the outputs of several segmentation algorithms often leads to improved performances [Menze et al., 2015]. One can note that this problem is related to Bayesian model averaging, which consists in making predictions according to a weighted combination of models instead of relying on a single one, thus reducing the risk of overconfidence [Hoeting et al., 1999].

Prior work has mainly focused on the fusion of binary masks. Majority voting is perhaps the most simple method and consists in choosing pixel-wise the most predominant

label among raters. A major limitation of this approach is the equal contribution of all raters to the consensus thus neglecting their potentially varying levels of performance. One of the most well known method proposed to address this issue is the STAPLE algorithm [Warfield et al., 2004]. It implements a weighting strategy based on the estimated level of performance of each expert. In this case, the raters' binary segmentations are described by Bernoulli distributions and an expectation-maximization (EM) scheme allows a consensus to be built and the raters' performances to be assessed at the same time. Spatial correlation between voxels is taken into account by the introduction of a Markov random field (MRF) prior over the consensus segmentation.

Among the known shortcomings of STAPLE, there is the constraint of having only global performance estimations of raters, and thus ignoring local variations [Commowick et al., 2012; Asman & Landman, 2012; Asman & Landman, 2011]. One proposed solution [Commowick et al., 2012] is to apply STAPLE in a sliding window fashion or to extend the performance parameters to the pixel level [Asman & Landman, 2012]. Another limitation is that STAPLE only considers binary masks as input and is thus agnostic to the image content and especially to the presence of large image gradients [Asman & Landman, 2013; Liu et al., 2013; Akhondi-Asl et al., 2014]. In [Liu et al., 2013], the authors proposed to include in the STAPLE approach simple appearance models, such as Gaussian distributions for the background and foreground, but this approach is only applicable to simple salient structures. Other extensions of STAPLE consider the case of missing data or repeated labels [Landman et al., 2012; Commowick & Warfield, 2010].

A first extension of the STAPLE algorithm for continuous inputs, which is the focus of this chapter, was proposed in [Warfield et al., 2008]. Raters' performances were captured by a set of biases and variances while assuming a Gaussian distribution for their continuous labels. This model was further studied in [Xing et al., 2016] and the authors demonstrated that to properly estimate rater bias, the introduction of a bias prior was required. An additional limitation of this model is the absence of a spatial prior for regularizing the consensus estimate. Furthermore, rater performances are not estimated locally but assumed to be global for the whole image, which was a limitation also shared by its binary counterpart, as noted above. Another model developed for probabilistic maps is PSTAPLE proposed by [Akhondi-Asl & Warfield, 2013]. This approach is closer to the binary STAPLE formalism than [Warfield et al., 2008] and also uses an MRF prior to regularize the consensus. However, raters performances are again estimated globally for the whole image.

In this chapter, we introduce a comprehensive probabilistic framework that addresses many shortcomings of approaches proposed in the literature for the fusion of continuous or categorical labels. Our baseline is the continuous STAPLE model introduced in [Warfield et al., 2008]. First, we propose replacing the Gaussian likelihood with heavy-tailed distributions to model the rater input maps. In this chapter, heavy-tailed distributions are broadly defined as distributions whose tails decline more slowly than the Gaussian distribution. Heavy-tailed distributions, unlike the Gaussian, are not very sensitive to outliers and, importantly, allow a spatial assessment of rater performances. Thus, image regions that differ greatly from the consensus segmentation will be considered as outliers and the contribution of that rater to the consensus will be reduced in the problematic area. In particular, the Laplace, Student's t and generalized double Pareto distributions are investigated. These distributions were used in prior works for their attractive robustness and sparsity-inducing properties. For instance, the Bayesian lasso that enables variable selection is based on the Laplace distribution [Park & Casella, 2008].

A robust Bayesian clustering approach was proposed in [Archambeau & Verleysen, 2007] using Student's t-distributions and a framework based on the generalized double Pareto distribution was developed for compressive sensing in [Sadeghigol et al., 2016]. In this chapter, we employ these distributions in a multivariate setting, which has never been done before for the generalized double Pareto distribution, to the best of our knowledge. In addition, we introduce a bias prior and take into account spatial correlation between voxels with a label smoothness prior, defined as a generalized linear model of spatially smooth kernels. We propose a common inference scheme based on variational calculus that allows the latent posterior distributions and the model parameters to be estimated in a data-driven fashion. Tractability is ensured for all heavy-tailed distributions by the use of scale mixture representations.

Last but not least, we address the unexplored issue of dissensus rather than consensus among raters. Indeed, fusing several probability maps into a single consensus map may not be meaningful when consistent patterns appear among raters. In [Langerak et al., 2010], the worse performing raters' masks were removed from the consensus estimation process at each iteration. In [Commowick & Warfield, 2009], a comparison framework for the raters' maps based on the continuous STAPLE parameters was developed. In the approach presented in this chapter, several consensuses are iteratively estimated through a technique similar to variational boosting [Miller et al., 2017] and clusters of raters are identified.

Finally, although our framework is particularly suitable for the fusion of continuous probability maps generated as is by segmentation algorithms, it can also be used for merging binary masks once they are transformed to the continuous domain using, for instance, signed distance maps [Pohl et al., 2007].

We summarize the main contributions of our work below:

- The classical Gaussian likelihood used in prior work is replaced by heavy-tailed distributions to model the input rater maps. This allows raters' performances to vary locally and their contributions to the consensus to be weighted differently depending on the region in the image.
- Heavy-tailed distributions are employed in a multivariate setting, which is novel for the generalized double Pareto distribution, to the best of our knowledge.
- Bias and spatial priors are introduced, allowing a proper bias estimation and a control over the smoothness of the consensus map.
- Tractability is ensured with a common variational inference scheme and scale mixture representations.
- The concept of a mixture of consensuses is introduced with a proper model and inference framework.

This chapter is built upon an earlier work of the authors [Audelan et al., 2020]. The initially proposed framework relying on a Student't-distribution is expanded with the introduction of two other heavy-tailed distributions, namely the Laplace and generalized double Pareto distributions. The relationships between these distributions is discussed and a common inference framework is proposed. Moreover, we also provide more extensive experiments and further analysis. In particular, we investigate for the mixture of consensuses a new application to raters clustering. The code used to perform the

experiments reported in this chapter is available in this repository: https://gitlab.inria.fr/epione/promfusion.

The rest of the chapter is organized as follows. Section 5.2 begins with the introduction of the robust probabilistic framework and the presentation of the heavy-tailed distributions investigated. Then, the common inference scheme based on variational calculus is developed, with details specific to each distribution. Section 5.3 explores the concept of a mixture of consensuses with a novel fusion algorithm similar to variational boosting. Finally, the last section gives qualitative and quantitative results on two datasets of prostate and lung nodule segmentations. We show that local variations in rater performance were successfully identified and that improved segmentation performances were obtained after fusing probability maps.

5.2 Robust estimate of consensus probability maps

5.2.1 Baseline probabilistic framework

The starting point of our work is the probabilistic framework proposed in [Warfield et al., 2008]. We are given as input a set of P probability maps $\mathbf{D}_n^p$, each map consisting of N categorical probability values in K classes, i.e. $\mathbf{D}_n^p \in S^{K-1} \in \mathbb{R}^K$ where S^{K-1} is the K unit simplex space such that $\sum_{k=1}^{K} \mathbf{D}_{nk}^p = 1$. P is the number of raters. In this chapter, a rater denotes either a human expert or a segmentation algorithm. Our objective is to estimate a consensus probability map $\mathbf{T}_n \in S^{K-1}$ over the input maps.

Each probability map is supposed to be derived from a consensus map through a random process. Let F be a link function $F(\mathbf{p}) \in \mathbb{R}^K$, where $\mathbf{p} \in S^{K-1}$, which maps probability S^{K-1} space into Euclidean space, and its inverse $F^{-1}(\mathbf{r})$ such that $F^{-1}(F(\mathbf{p})) = \mathbf{p}$. We write $\tilde{\mathbf{D}}_n^p = F(\mathbf{D}_n^p)$ and $\tilde{\mathbf{T}}_n = F(\mathbf{T}_n)$. In this chapter, we follow [Pohl et al., 2007] and consider the logit function and its inverse as link functions. For instance, we have for $K = 2$:

$$F((\mathbf{p}_1, \mathbf{p}_2)^T) = \left(\log \frac{\mathbf{p}_1}{1 - \mathbf{p}_1}, \log \frac{\mathbf{p}_2}{1 - \mathbf{p}_2} \right)^T , \tag{5.1}$$

$$F^{-1}((\mathbf{r}_1, \mathbf{r}_2)^T) = \left(\frac{\sigma(\mathbf{r}_1)}{\sigma(\mathbf{r}_1) + \sigma(\mathbf{r}_2)}, \frac{\sigma(\mathbf{r}_2)}{\sigma(\mathbf{r}_1) + \sigma(\mathbf{r}_2)} \right)^T , \tag{5.2}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function.

Our baseline model follows prior works [Warfield et al., 2008; Xing et al., 2016] and assumes that the observed rater probability maps $\tilde{\mathbf{D}}^p$ are Gaussian distributed with a mean given by the consensus plus a rater bias:

$$p(\tilde{\mathbf{D}}_n^p | \tilde{\mathbf{T}}_n^p, \mathbf{b}_p, \mathbf{\Sigma}_p) = \mathcal{N}\left(\tilde{\mathbf{D}}_n^p; \tilde{\mathbf{T}}_n + \mathbf{b}_p, \mathbf{\Sigma}_p\right) . \tag{5.3}$$

The rater bias $\mathbf{b}_p$ and variance $\mathbf{\Sigma}_p$ do not depend on the location in the image. Together, they characterize the rater performance at the whole image level, large biases and variances being associated with poor performances.

In [Xing et al., 2016], the authors demonstrated that the absence of a bias prior leads to an indeterminate estimation. Therefore, we define a zero mean Gaussian prior over the bias, with precision β:

$$p(\mathbf{b}_p|\beta) = \mathcal{N}(\mathbf{b}; 0, \beta^{-1}\mathbf{I}_K)\,. \tag{5.4}$$

Moreover, spatial smoothness is generally considered as a desirable characteristic of segmentation maps. In the binary case [Warfield et al., 2004], a Markov random field (MRF) prior was introduced to allow a connectivity-based regularization of the discrete consensus map. Spatial consistency was also enforced through an MRF prior in PSTAPLE [Akhondi-Asl & Warfield, 2013], which is another approach extending STAPLE to continuous inputs. The main limitation of MRF priors is the impossibility of a data-driven estimation of the MRF hyperparameter β controlling the level of regularization. Because inference cannot be done in closed-form, it has to be set manually. In the context of our probabilistic framework, prior works [Warfield et al., 2008; Xing et al., 2016] did not include any smoothness prior.

In our model, spatial regularity of the consensus map is enforced by a smoothness prior defined as a generalized linear model of a set of L spatially smooth functions $\{\Phi_l(\mathbf{x})\}$, whose hyperparameters can be estimated. Let $\mathbf{x}_n \in \mathbb{R}^D$ be the position of voxel n, where D is the image dimension. Then the prior on the variables $\tilde{\mathbf{T}}_n$ is defined as:

$$p(\tilde{\mathbf{T}}_n|\mathbf{W}_l) = \mathcal{N}\left(\tilde{\mathbf{T}}_n; \sum_{l=1}^{L} \Phi_l(\mathbf{x}_n)\mathbf{W}_l; \mathbf{\Sigma}_T\mathbf{I}_K\right)\,, \tag{5.5}$$

where $\mathbf{W}_l$ are vectors of size K and where $\mathbf{\Sigma}_T \in \mathbb{R}^+$ is the prior variance. For computational convenience, we write the prior using $\mathbf{W}_k \in \mathbb{R}^L$, such that $p(\tilde{\mathbf{T}}_{nk}|\mathbf{W}_k) = \mathcal{N}(\tilde{\mathbf{T}}_{nk}; \mathbf{W}_k^T\mathbf{\Phi}_n, \mathbf{\Sigma}_T)$ where $\mathbf{\Phi}_n^T = [\Phi_1(\mathbf{x}_n), \cdots, \Phi_L(\mathbf{x}_n)]$. The weights $\mathbf{W}_k$ are placed in a weight matrix $\mathbf{W} \in \mathbb{R}^{K\times L}$ such that we can write more compactly:

$$p(\tilde{\mathbf{T}}_n|\mathbf{W}) = \mathcal{N}(\tilde{\mathbf{T}}_n; \mathbf{W}\mathbf{\Phi}_n; \mathbf{\Sigma}_T\mathbf{I}_K) \tag{5.6}$$

To obtain a robust description, the weights $\mathbf{W}_k$ are equipped with a zero mean Gaussian prior and precision α:

$$p(\mathbf{W}_k|\alpha) = \mathcal{N}(\mathbf{W}_k; 0, \alpha^{-1}\mathbf{I}_L)\,. \tag{5.7}$$

The spatial prior will be denoted by GLSP (Generalized Linear Spatial Prior) in the remainder of the chapter. The graphical model of this baseline framework is shown in Fig. 5.1.

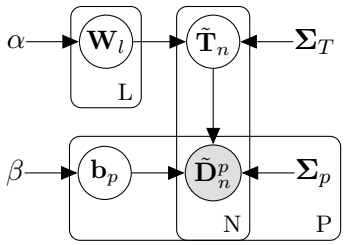


Fig. 5.1: Graphical representation of the baseline model with a Gaussian likelihood.

5.2.2 Heavy-tailed distributions and scale mixture representation

The main limitation of the baseline model presented in the last section is the global estimation of rater performances, thus neglecting local variations. In this chapter, we address this issue by replacing the Gaussian distribution with heavy-tailed likelihoods. More specifically, we propose the Laplace, Student's t and generalized double Pareto (GDP) distributions as substitutes.

A relationship between these distributions can be established by first introducing the power exponential distribution, also known as generalized Gaussian distribution [Pascal et al., 2013; Gómez et al., 1998]. The density function of a multivariate power exponential distribution is written:

$$\mathrm{PE}(x;\tau,M,\theta) = |M|^{-\frac{1}{2}} h_{\theta,\tau}((x-\mu)^T M^{-1}(x-\mu)) \;, \tag{5.8}$$

for $x \in \mathbb{R}^K$ where M is a $K \times K$ covariance matrix, $\theta > 0$, $\tau > 0$, and

$$h_{\theta,\tau}(y) = \frac{\theta\Gamma\left(\frac{K}{2}\right)}{\pi^{\frac{K}{2}}\Gamma\left(\frac{K}{2\theta}\right)} \tau^{\frac{K}{2\theta}} \exp\left(-\tau y^{\theta}\right) \;. \tag{5.9}$$

Power exponential scale mixtures are distributions that can be represented in a hierarchical fashion using a scale mixture as follows:

$$p_X(x) = \int_\tau p_{X|\tau}(x) p_\tau(\tau) d\tau = \int_\tau \mathrm{PE}(x;\tau,M,\theta) p_\tau(\tau) d\tau \;. \tag{5.10}$$

Depending on the choice of parameter θ and mixing density $p_\tau(\tau)$, various distributions can be obtained. In this chapter, we consider the case where the mixing density is a Gamma distribution $p_\tau(\tau) = \mathrm{Ga}(\tau;\nu,\nu)$ with shape and scale parameter $\nu > 0$:

$$\mathrm{Ga}(x;\nu,\nu) = \frac{\nu^\nu}{\Gamma(\nu)} x^{\nu-1} \exp\left(-\nu x\right) \;. \tag{5.11}$$

Then, the resulting distribution $p_X(x)$ obtained after marginalization of τ is a generalized t distribution [Giri, 2016] whose density function is given by:

$$p_X(x) = \frac{\theta\Gamma\left(\frac{K}{2}\right)\nu^\nu}{\pi^{\frac{K}{2}} B\left(\nu,\frac{K}{2\theta}\right)} |M|^{-\frac{1}{2}} \times \frac{1}{\left(\nu + ((x-\mu)^T M^{-1}(x-\mu))^{\theta}\right)^{\nu+\frac{K}{2\theta}}} \;, \tag{5.12}$$

where $\Gamma(x)$ and $B(a,b)$ are the Gamma and Beta functions, respectively [Arslan, 2004]. Depending on the values of θ and ν, different situations can arise [Giri, 2016]:

- If $\theta = 1$, we get a multivariate Student's t-distribution. Moreover, if $\nu \longrightarrow \infty$ then we recover the multivariate Gaussian.
- If $\theta = \frac{1}{2}$, we obtain a multivariate generalized double Pareto distribution. Moreover, if $\nu \longrightarrow \infty$ then we recover the multivariate Laplace distribution.

Together, the θ and ν parameters control the shape of the distribution tails. Large parameters values lead to thinner tails while smaller values lead to heavier tails [McDonald & Newey, 1988]. Fig. 5.2 shows the four distributions and compares the tails for different

parameter values. The Laplace distribution spikes at zero and has fatter tails than the baseline Gaussian. The Student's t and GDP distributions have with the parameter ν a supplementary degree of freedom in comparison with the Gaussian and Laplace distributions, allowing the level of robustness to be adapted.

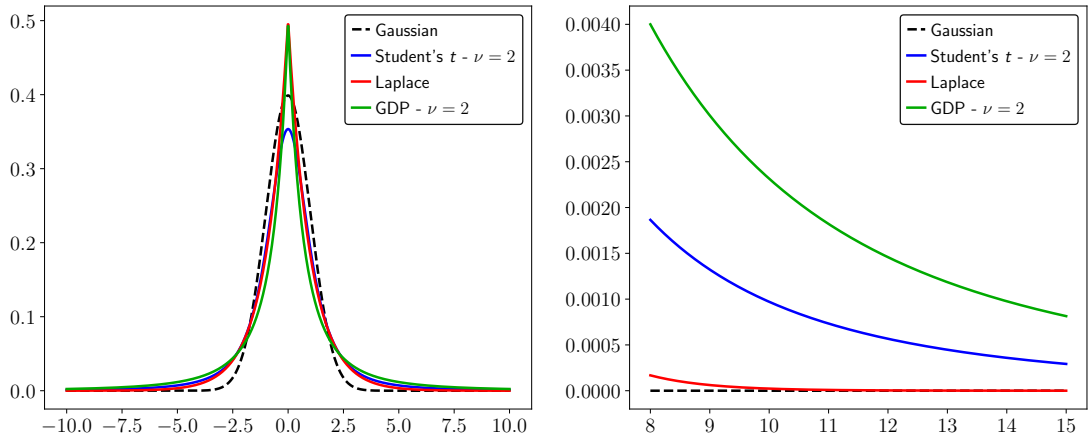


Fig. 5.2: Density plot of the zero-mean heavy-tailed distributions (5.2a), with a focus on the tail behaviors (5.2b).

Interestingly, the three heavy-tailed distributions can all be written as Gaussian scale mixtures, namely $p_X(x) = \int_\tau \mathcal{N}\left(x; \mu, \frac{M}{\tau}\right) p_\tau(\tau) d\tau$. This re-writing is attractive for 2 reasons. First, it enables the model tractability by leading to closed-form analytical solutions within a variational Bayes framework. Second, it introduces a new variable, the scale factor τ, that can be used to make the rater variance dependent on the location in the image.

The derivation of the scale mixture for the Student's t is straightforward as the power exponential distribution of Eq. 5.10 amounts to a Gaussian for $\theta = 1$. The same equation for $\theta = \frac{1}{2}$ corresponds to a Laplace scale mixture. Yet, [Gómez-Sánchez-Manzano et al., 2008] showed that for any $\theta \in]0, 1]$, the power exponential can be written as a Gaussian scale mixture. However, as the mixing densities involve stable distributions, they cannot generally be written analytically, except for a few cases and in particular for $\theta = \frac{1}{2}$. The Laplace and generalized double Pareto distributions can thus be written as Gaussian scale mixtures, with an additional level of hierarchy for the latter.

Tab. 5.1 summarizes how the rater input map distributions, $p(\tilde{\mathbf{D}}_n^p)$, are written as scale mixtures after replacement of the Gaussian with the heavy-tailed distributions. The corresponding graphical models are presented in Fig. 5.3. The scale factors $\{\tau_n^p\} \in \mathbb{R}^{+N}$ are additional latent variables not present in the Gaussian model, that separately weight each data point $\tilde{\mathbf{D}}_n^p$, allowing local variations in the performance of rater p to be taken into account. The degree of freedom ν_p^{-1} characterizes the number of data outliers that it is necessary to discard in the estimation of the consensus, i.e., a small degree of freedom ν_p indicates that rater p contributes a lot of outliers.

One can note that prior knowledge could be incorporated over the model parameters α and β by introducing, for example, Gamma hyperpriors. However, this is not the choice made in this chapter, where we consider a simpler situation with uniform priors.

Tab. 5.1: Heavy-tailed distributions and scale mixture representations. $S_n^p = \{\tau_n^p\}$ for the Student's t and Laplace distributions, and $S_n^p = \{\tau_n^p, z_n^p\}$ for the generalized double Pareto distribution.

Likelihood	Parameters	$p(\tilde{\mathbf{D}}_n^p \vert \tilde{\mathbf{T}}_n, \mathbf{b}_p, \mathbf{\Sigma}_p, S_n^p)$
Student's t	$\theta = 1$ $\nu > 0$	$\int_{\tau_n^p} \mathcal{N}\left(\tilde{\mathbf{D}}_n^p; \tilde{\mathbf{T}}_n + \mathbf{b}_p, \frac{\mathbf{\Sigma}_p}{\tau_n^p}\right) \mathrm{Ga}\left(\tau_n^p; \frac{\nu_p}{2}, \frac{\nu_p}{2}\right) \, d\tau_n^p$
Laplace	$\theta = \frac{1}{2}$ $\nu \longrightarrow \infty$	$\int_{\tau_n^p} \mathcal{N}\left(\tilde{\mathbf{D}}_n^p; \tilde{\mathbf{T}}_n + \mathbf{b}_p, \frac{\mathbf{\Sigma}_p}{\tau_n^p}\right) \mathrm{InvGa}\left(\tau_n^p; \frac{K+1}{2}, \frac{1}{8}\right) d\tau_n^p$
GPD	$\theta = \frac{1}{2}$ $\nu > 0$	$\int_{\tau_n^p}\int_{z_n^p} \mathcal{N}\left(\tilde{\mathbf{D}}_n^p; \tilde{\mathbf{T}}_n + \mathbf{b}_p, \frac{\mathbf{\Sigma}_p}{\tau_n^p}\right) \mathrm{InvGa}\left(\tau_n^p; \frac{K+1}{2}, \frac{(z_n^p)^2}{2}\right) \mathrm{Ga}\left(z_n^p; \nu_p, \nu_p\right) dz_n^p d\tau_n^p$

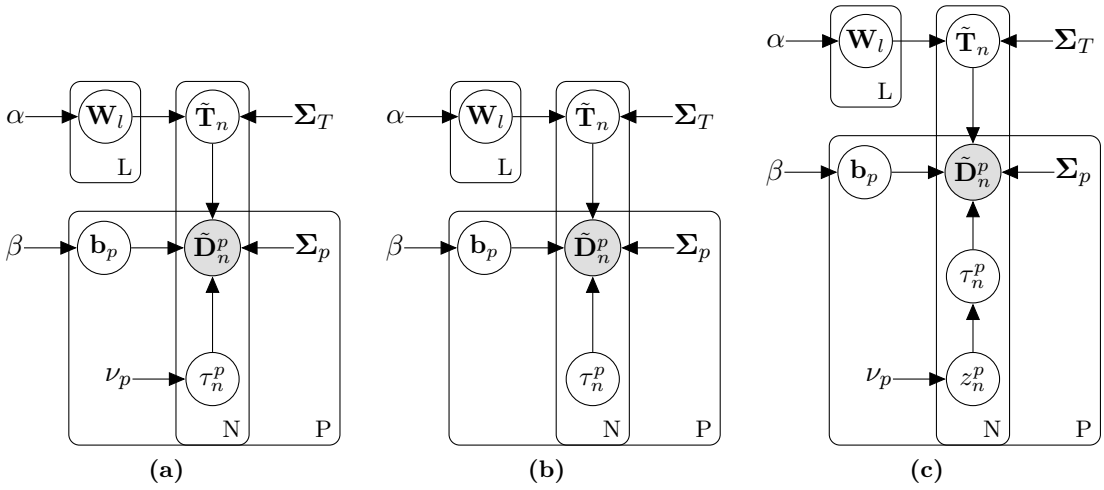


Fig. 5.3: Graphical models of the probabilistic framework with a Student's t-distribution (5.3a), a Laplace distribution (5.3b) and a generalized double Pareto distribution (5.3c).

5.2.3 Model inference

To estimate the consensus, previous works used an EM algorithm. However, this approach does not lead to closed-form solutions after replacing the Gaussian with heavy-tailed distributions. Instead, we propose a common inference framework based on variational calculus (a.k.a. variational Bayes) allowing the true posterior distribution $p(U|\tilde{\mathbf{D}})$ of the model variables $U = \{\tilde{\mathbf{T}}, \mathbf{b}, S, \mathbf{W}\}$ to be approximated by a chosen family of distributions $q(U)$. We recall that $S = \{\tau\}$ for the Student's t and Laplace distributions, and $S = \{\tau, z\}$ for the generalized double Pareto distribution.

The objective is to maximize the marginal log likelihood of the data by minimizing the Kullback-Leibler divergence between the true posterior $p(U|\tilde{\mathbf{D}})$ and the approximation $q(U)$, or equivalently by maximizing the lower bound $\mathcal{L}(q)$:

$$\log p(\tilde{\mathbf{D}}) = \underbrace{\int_U q(U) \log \frac{p(\tilde{\mathbf{D}}, U)}{q(U)}}_{\mathcal{L}(q)} + \underbrace{\mathrm{KL}[q(U)||p(U|\tilde{\mathbf{D}})]}_{\geq 0} . \tag{5.13}$$

Furthermore, we assume a mean field approximation leading to a factorization of the posterior approximation as follows:

$$q(U) = q(\tilde{\mathbf{T}})q(\mathbf{b})q(S)q(\mathbf{W}) . \tag{5.14}$$

The lower bound can be re-written as:

$$\begin{aligned} \log p(\tilde{\mathbf{D}}) \geq \mathcal{L}(q) = \int_{\tilde{\mathbf{T}}} \int_{\mathbf{b}} \int_S \int_{\mathbf{W}} & q(\tilde{\mathbf{T}})q(\mathbf{b})q(S)q(\mathbf{W}) \\ & \log \frac{p(\tilde{\mathbf{D}}, \tilde{\mathbf{T}}, \mathbf{b}, S, \mathbf{W})}{q(\tilde{\mathbf{T}})q(\mathbf{b})q(S)q(\mathbf{W})} d\tilde{\mathbf{T}}\, d\mathbf{b}\, dS\, d\mathbf{W} . \end{aligned} \tag{5.15}$$

If q_i denotes any of the factors in Eq. 5.14 and q_{-i} the product of the remaining factors, we know by variational calculus that the distribution q_i^* maximizing Eq. 5.15 has the form:

$$\log q_i^* = \mathbb{E}_{q_{-i}}[\log p(\tilde{\mathbf{D}}, U)] + cst , \tag{5.16}$$

when fixing the other distributions q_{-i}.

This results leads to an iterative algorithm where the lower bound is optimized with respect to each approximate distribution q_i in turn. We present in the following sections the main results for each posterior distribution approximation. Details about the derivations can be found in appendix C.1 and the values of some expectations are compiled in appendix C.3.

Consensus posterior approximation.

Using Eq. 5.16, the consensus posterior approximation is found to be Gaussian distributed $\mathcal{N}(\tilde{\mathbf{T}}_n; \mu_{\tilde{\mathbf{T}}_n}, \mathbf{\Sigma}_{\tilde{\mathbf{T}}_n})$, with parameters given by:

$$\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n} = \left[\sum_{p=1}^{P} \mathbb{E}[\tau_{np}] \mathbf{\Sigma}_p^{-1} + \Sigma_T^{-1} \mathbf{I}_K \right]^{-1} , \tag{5.17}$$

$$\mu_{\tilde{\mathbf{T}}_n} = \mathbf{\Sigma}_{\tilde{\mathbf{T}}_n} \left[\sum_{p=1}^{P} \mathbb{E}[\tau_{np}] \mathbf{\Sigma}_p^{-1} \left(\tilde{\mathbf{D}}_n^p - \mathbb{E}[\mathbf{b}_p] \right) + \Sigma_T^{-1} \mathbb{E}[\mathbf{W}] \mathbf{\Phi}_n \right] . \tag{5.18}$$

With a Gaussian likelihood, the consensus mean vector at voxel n was given by $\mu_{\tilde{\mathbf{T}}_n} = \mathbf{\Sigma}_{\tilde{\mathbf{T}}_n} \left[\sum_{p=1}^{P} \mathbf{\Sigma}_p^{-1} \left(\tilde{\mathbf{D}}_n^p - \mathbb{E}[\mathbf{b}_p] \right) + \Sigma_T^{-1} \mathbb{E}[\mathbf{W}] \mathbf{\Phi}_n \right]$. Thus, the consensus is now computed as a weighted mean of the raters' values corrected with the bias, where the weights vary spatially through the variable τ according to the raters' local performances.

Rater bias posterior approximation.

The posterior approximation of the rater bias is also a Gaussian distribution $\mathcal{N}(\mathbf{b}_p; \mu_{\mathbf{b}_p}, \mathbf{\Sigma}_{\mathbf{b}_p})$, whose parameters are given below:

$$\mathbf{\Sigma}_{\mathbf{b}_p} = \left[\beta \mathbf{I}_K + \sum_{n=1}^{N} \mathbb{E}[\tau_{np}] \mathbf{\Sigma}_p^{-1}\right]^{-1} , \tag{5.19}$$

$$\mu_{\mathbf{b}_p} = \mathbf{\Sigma}_{\mathbf{b}_p} \sum_{n=1}^{N} \mathbb{E}[\tau_{np}] \mathbf{\Sigma}_p^{-1} \left(\tilde{\mathbf{D}}_n^p - \mathbb{E}[\tilde{\mathbf{T}}_n]\right) . \tag{5.20}$$

Posterior approximations of the scale variables.

The scale mixture representation introduced supplementary latent variables, the scale factor τ common to the three distributions and, in addition, z for the generalized double Pareto distribution.

Applying Eq. 5.16 for the first scale factor leads to a Gamma distribution for the Student's t framework and to an inverse Gaussian distribution for the other two. Formula are given in Tab 5.2.

Tab. 5.2: Posterior approximation of the scale factor τ depending on the chosen likelihood. E is given by $E = \mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)^T \mathbf{\Sigma}_p^{-1} (\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)]$.

Likelihood	$q(\tau_n^p)$	Density	Parameters
Student's t	$\mathrm{Ga}(\tau_n^p; a_{np}, b_{np})$	$\frac{x^{a_{np}-1} b_{np}^{a_{np}}}{\Gamma(a_{np})} \exp(-b_{np}x)$	$a_{np} = \frac{\nu_p + K}{2},\ b_{np} = \frac{\nu_p}{2} + \frac{E}{2}$
Laplace	$\mathcal{IG}(\tau_n^p; \mu_{np}, \lambda_{np})$	$\sqrt{\frac{\lambda_{np}}{2\pi x^3}} \exp\left(-\frac{\lambda_{np}(x-\mu_{np})^2}{2\mu_{np}^2 x}\right)$	$\mu_{np} = \frac{1}{2\sqrt{E}},\ \lambda_{np} = \frac{1}{4}$
GDP	$\mathcal{IG}(\tau_n^p; \mu_{np}, \lambda_{np})$	$\sqrt{\frac{\lambda_{np}}{2\pi x^3}} \exp\left(-\frac{\lambda_{np}(x-\mu_{np})^2}{2\mu_{np}^2 x}\right)$	$\mu_{np} = \sqrt{\frac{\mathbb{E}[(z_n^p)^2]}{E}},\ \lambda_{np} = \mathbb{E}[(z_n^p)^2]$

The GDP model has a supplementary level of hierarchy with the other scale variable z. Eq. 5.16 leads to the following equation for $q^*(z_n^p)$:

$$q^*(z_n^p) = \frac{(\mathcal{T}_n^p)^{\frac{K+\nu_p+1}{2}} (z_n^p)^{K+\nu_p} \exp\left(-\nu_p z_n^p - \frac{(z_n^p)^2}{2}\mathcal{T}_n^p\right)}{\Gamma(K+\nu_p+1) \exp(\frac{\nu_p^2}{4\mathcal{T}_n^p}) D_{-K-\nu_p-1}\left(\frac{\nu_p}{\sqrt{\mathcal{T}_n^p}}\right)} , \tag{5.21}$$

where $\mathcal{T}_n^p$ stands for $\mathbb{E}\left[\frac{1}{\tau_n^p}\right] = \frac{1}{\mu_{np}} + \frac{1}{\lambda_{np}}$ and D_ν is the parabolic cylinder function of order $\nu \in \mathbb{R}$. The expectations $\mathbb{E}[z_n^p]$ and $\mathbb{E}[(z_n^p)^2]$ can be computed and are given in appendix C.1.

Spatial regularization variable.

We now present the posterior approximation for the variable $\mathbf{W}_k$ which controls the smoothness of the k-th consensus map. $q^*(\mathbf{W}_k)$ is a Gaussian distribution $\mathcal{N}(\mathbf{W}_k, \mu_{\mathbf{W}_k}, \mathbf{\Sigma}_{\mathbf{W}_k})$ whose parameters are:

$$\mathbf{\Sigma}_{\mathbf{W}_k} = \left[\mathbf{\Sigma}_T^{-1} \left(\sum_{n=1}^{N} \mathbf{\Phi}_n \mathbf{\Phi}_n^T \right) + \alpha \mathbf{I}_L \right]^{-1}, \tag{5.22}$$

$$\mu_{\mathbf{W}_k} = \mathbf{\Sigma}_{\mathbf{W}_k} \left[\sum_{n=1}^{N} \mathbf{\Phi}_n \Sigma_T^{-1} \mathbb{E}[\tilde{\mathbf{T}}_{nk}] \right]. \tag{5.23}$$

Update of the model parameters.

Finally, a data-driven estimation of the model parameters can be performed. The parameters in question are α, which controls the extent of the spatial regularization, $\mathbf{\Sigma}_T$, the covariance of the consensus prior, β, the precision of the prior defined over the rater bias, $\mathbf{\Sigma}_p$, the rater variance and lastly ν_p, the degree of freedom of the Student's t and GDP distributions.

We assume that the posterior approximation of these parameters is a Dirac distribution. Applying Eq. 5.16 and taking the derivatives, we obtain the following update formula:

$$\alpha = \frac{LK}{\sum_{k=1}^{K} \mu_{\mathbf{W}_k}^T \mu_{\mathbf{W}_k} + \mathrm{Tr}(\mathbf{\Sigma}_{\mathbf{W}_k})}, \tag{5.24}$$

$$\mathbf{\Sigma}_T = \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} (\mu_{\tilde{\mathbf{T}}_{nk}} - \mu_{\mathbf{W}_k}^T \mathbf{\Phi}_n)^2 + \mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nk}} + \mathrm{Tr}(\mathbf{\Phi}_n \mathbf{\Phi}_n^T \mathbf{\Sigma}_{\mathbf{W}_k})}{NK}, \tag{5.25}$$

$$\beta = \frac{KP}{\sum_{p=1}^{P} \mu_{\mathbf{b}_p}^T \mu_{\mathbf{b}_p} + \mathrm{Tr}(\mathbf{\Sigma}_{\mathbf{b}_p})}, \tag{5.26}$$

$$\begin{aligned} \mathbf{\Sigma}_p = \frac{1}{N} \sum_{n=1}^{N} \Big(& (\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p}) \mathbb{E}[\tau_n^p] (\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p})^T \\ & + \mathbb{E}[\tau_n^p] (\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n} + \mathbf{\Sigma}_{\mathbf{b}_p}) \Big). \end{aligned} \tag{5.27}$$

Finally, finding the mode of $q^*(\nu_p)$ leads to the following equation when the likelihood is a Student's t-distribution:

$$\sum_{n=1}^{N} -\psi\left(\frac{\nu_p}{2}\right) + \log\frac{\nu_p}{2} + 1 + \mathbb{E}[\log \tau_n^p] - \mathbb{E}[\tau_n^p] = 0, \tag{5.28}$$

with ψ being the digamma function. In practice, the ν_p are updated by solving the equation numerically. A similar approach could be implemented for optimizing the degree of freedom of the GDP distribution. However in practice, the numerical optimization is very unstable and we decided to set this parameter manually in the remainder of the chapter.

5.3 Mixture of consensuses

5.3.1 Probabilistic framework

We also investigate the issue of dissensus rather than consensus among raters and propose a novel probabilistic framework that allows a mixture of consensuses to be estimated.

We now assume that the rater maps are derived not from a single map but from M consensus maps. We introduce for each rater a new binary latent variable $\mathbf{Z}_{pm} \in \{0, 1\}$, $\sum_m \mathbf{Z}_{pm} = 1$, specifying from which consensus a rater map is generated. The associated component prior is given by the mixing coefficients π_m such that $p(\mathbf{Z}_{pm} = 1) = \pi_m$. Moreover, we consider a simpler model than in the previous section, by removing the rater bias and assuming that the rater input probability maps are Gaussian distributed, i.e.:

$$p(\tilde{\mathbf{D}}^p|\tilde{\mathbf{T}}) = \prod_{m=1}^{M} \mathcal{N}(\tilde{\mathbf{D}}^p; \tilde{\mathbf{T}}_m, \mathbf{\Sigma}_p)^{\mathbf{Z}_{pm}} . \tag{5.29}$$

The graphical model of the mixture of consensuses is presented in Fig. 5.4.

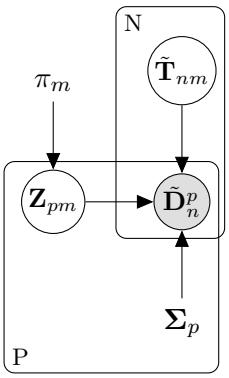


Fig. 5.4: Graphical model of the mixture of consensuses

5.3.2 Model inference

As for the robust probabilistic framework, we use variational inference to infer the consensus and model parameters. A naive solution would compute the posterior component probabilities, r_{pm} (a.k.a. the responsibilities), as a classical Gaussian mixture clustering problem with multivariate Gaussians of dimension N, thus leading to dubious results due to the curse of dimensionality (high dimension, few samples).

Instead, we propose to first reduce the dimension of each rater input map by applying a principal component analysis (PCA) and then to cluster the maps in this low-dimensional space. The resulting consensus maps are obtained by applying the inverse mapping from the component weights to the original space.

We assume again a mean field approximation implying that the approximation of the posterior factorizes as $q(U) = q(\mathbf{Z})q(\tilde{\mathbf{T}})$ with $U = \{\mathbf{Z}, \tilde{\mathbf{T}}\}$. The optimal approximate distribution q_i^* maximizing the lower bound is given as before by Eq. 5.16. The following sections present the main results for each variational update; details of the derivations can be found in appendix C.1.

Label posterior approximation.

The variable $\mathbf{Z}$ indicates from which consensus each rater input map is generated. Eq. 5.16 applied to $q(Z_p)$ leads to a categorical distribution of parameters r_{pm}, with $r_{pm} = \rho_{pm} / \sum_m \rho_{pm}$ for $1 \leq m \leq M$, and:

$$\begin{aligned} \log \rho_{pm} = \log \pi_m + \sum_{n=1}^{N} \Bigg(& -\frac{K}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{\Sigma}_p| \\ & - \frac{1}{2} \mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})^T \mathbf{\Sigma}_p^{-1} (\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})] \Bigg) . \end{aligned} \tag{5.30}$$

Consensus posterior approximation.

There is no longer a unique consensus but M consensuses to estimate. The approximate posterior distribution for each of them is a Gaussian distribution $\mathcal{N}(\tilde{\mathbf{T}}_{nm}; \mu_{\tilde{\mathbf{T}}_{nm}}, \mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}})$, whose parameters are written below:

$$\mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}} = \left[\sum_{p=1}^{P} r_{pm} \mathbf{\Sigma}_p^{-1} \right]^{-1} , \tag{5.31}$$

$$\mu_{\tilde{\mathbf{T}}_{nm}} = \mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}} \sum_{p=1}^{P} r_{pm} \mathbf{\Sigma}_p^{-1} \tilde{\mathbf{D}}_n^p . \tag{5.32}$$

The raters contributions to the consensus m are now weighted by the responsibilities r_{pm}, i.e, the posterior probabilities of being generated from the consensus in question for each rater.

Update of the model parameters.

The model parameters are the mixing coefficients π_m and the rater variance $\mathbf{\Sigma}_p$. The former is updated with the following formula:

$$\pi_m = \frac{\sum_{p=1}^{P} r_{pm}}{P} , \tag{5.33}$$

and the latter according to:

$$\mathbf{\Sigma}_p = \frac{1}{N} \left(\sum_{n=1}^{N} \sum_{m=1}^{M} r_{pm} \left((\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{nm}})(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})^T + \mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}} \right) \right) . \tag{5.34}$$

The inference has been found experimentally to be very sensitive to the initial values. To increase its stability, we follow an incremental scheme inspired by variational boosting [Miller et al., 2017]. We introduce one consensus map at a time and the distribution parameters of components included in the previous iterations are not updated. Initialization is performed at each iteration by summing the absolute value of the residuals $\text{res}_p = \sum_{n,m} |\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm}|$ and setting the responsibility for the new component to $\frac{\text{res}_p}{\sum_p \text{res}_\text{p}}$ for rater p. Other responsibilities are uniformly initialized such that $\sum_m r_{pm} = 1$. In practice, the algorithm is stopped when no rater is added to

Algorithm 2: Mixture of consensuses

Inputs:

- $\tilde{\mathbf{D}}$ `// raters continuous segmentation maps`
- M_{cons} `// maximum number of consensuses`

$\tilde{\mathbf{D}} = \text{PCA}(\tilde{\mathbf{D}})$ `// dimensionality reduction`
$m = 0$ `// current number of consensuses in the model`
while $m < M_{cons}$ **do**
  $m \longleftarrow m + 1$
  while *not converged* **do**
    for $1 \leq i \leq m$, $1 \leq p \leq P$ **do**
      Estimate r_{pi} and π_i from Eq. 5.30 and Eq. 5.33
      Estimate $\mathbf{\Sigma}_p$ from Eq 5.34
    end
    for $1 \leq n \leq N$, $i = m$ **do**
      `// distribution parameters of components already in the model at` $m-1$ `are not updated`
      Update $\mathbf{\Sigma}_{\tilde{\mathbf{T}}_{ni}}$ and $\mu_{\tilde{\mathbf{T}}_{ni}}$ from Eq. 5.31 and Eq. 5.32
    end
  end
  if $\pi_m < 10^{-10}$ **then**
    $m = M_{\text{cons}}$ `// stop when the new component is empty`
  end
end
$\mu_{\tilde{\mathbf{T}}} \longleftarrow \text{PCA}^{-1}(\mu_{\tilde{\mathbf{T}}})$ `// return to the original space`
return $\mu_{\tilde{\mathbf{T}}}$

the newly introduced component after convergence. The sketch of the approach is summarized in Alg. 2.

5.4 Results

5.4.1 Material

We investigate our approach on prostate and nodule segmentations. Two types of experiments were conducted, depending on the nature of the segmentations used as input.

In the first case, we used binary segmentations drawn by medical experts as inputs. The binary masks were first transformed into probabilistic segmentations computed as the sigmoid of a Euclidean signed distance map, whose 0 level corresponds to the segmentation boundaries. The sigmoid function is defined as $\sigma(x) = 1/(1 + \exp(-\lambda_s x))$, where λ_s controls the slope of the transition between regions. Small lambda values are associated with increased uncertainty along the segmentation border.

In the second case, the inputs were continuous segmentations produced by several neural networks, trained beforehand by cross-validation on an independent training set. The consensus estimated between the neural networks was then compared to a

surrogate ground truth defined as a majority vote of the medical experts' delineations. All networks used in this chapter have a classical U-net architecture [Ronneberger et al., 2015].

The prostate dataset is a private collection of 40 MRI exams performed at 3 tesla (SIGNA^{TM} Architect, GE Healthcare, Chicago, IL and MAGNETOM^{TM} Skyra, Siemens Healthcare, Erlangen, Germany). All MRI protocols included 3D T2 weighted images with 0.5 mm to 1.0 mm slice thickness. The in-plane pixel size ranges from 0.4 mm to 0.8 mm. The dataset includes manual prostate delineations from 7 radiologists, whose levels of experience are dissimilar: three are considered as experts, two have an intermediary level, and the remaining two are junior radiologists with less experience. This dataset, with binary segmentations, will be denoted latter as ProstateBin.

Moreover, 5 neural networks were trained by 5-fold cross-validation on a subset of 98 3D T2 weighted images selected from the publicly available SPIE-AAPM-NCI PROSTATEx dataset [Litjens et al., 2014], and for which [Meyer et al., 2019] released ground truth segmentations made by an expert urologist. The performances of the networks were then evaluated on 7 unseen test scans extracted from the private dataset of 40 images described above. This set composed of 7 images and of the associated predictions of the 5 neural networks, will be referred to as ProstateNet.

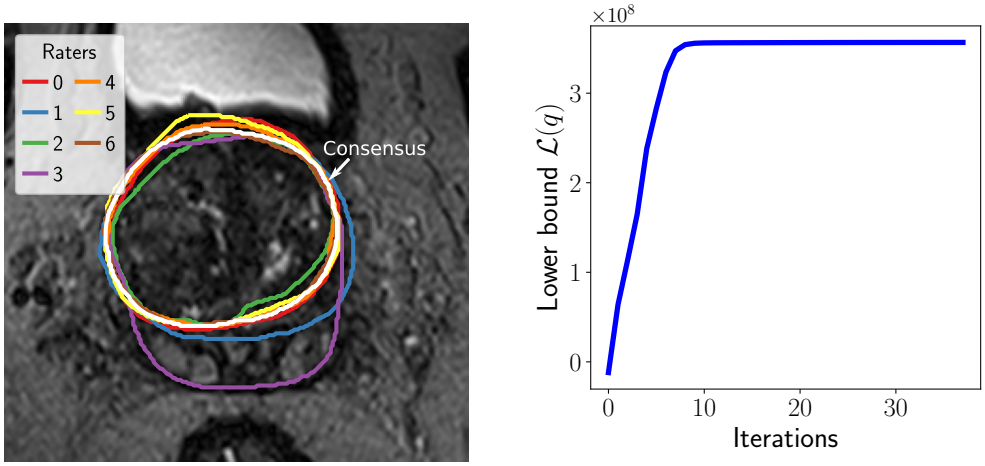
The nodule dataset is the publicly available LIDC-IDRI database of lung CT scans [Armato III et al., 2011]. It contains nodule delineations drawn by 4 radiologists. The raw CT images were re-sampled in a pre-processing step to obtain a common spatial resolution of 1 mm in all directions. A first set was constituted by considering the 20 largest nodules annotated by all radiologists. This set, containing 20 lesions and binary segmentations, will be denoted as NoduleBin in the remainder of the chapter. The LIDC-IDRI dataset was furthermore separated into a training and a testing set. The former was used to train 9 neural networks by 9-fold cross validation. The networks were then evaluated on the 34 nodules of the test set having a 10 mm minimum diameter. The set composed of the 34 test cases and the associated networks predictions will be referred to as NoduleNet in the remainder.

Tab. 5.3 summarizes the characteristics of the datasets used in the experiments. All results reported in this chapter were obtained in 3D. The size of the inputs depends on the dataset. For the experiments on the nodule datasets, we used a cube of size $48 \times 48 \times 48$ centered at each nodule location. Computations were performed on the entire image for the prostate datasets. The typical image size in the prostate datasets was $160 \times 500 \times 500$.

Tab. 5.3: Characteristics of the datasets used for the experiments. (MV: majority vote.)

	ProstateBin	ProstateNet	NoduleBin	NoduleNet
# of cases	40	7	20	34
# of experts	7	5	4	10
Expert category	Radiologists	Neural networks	Radiologists	Neural networks
Segmentation type	Binary	Continuous	Binary	Continuous
Surrogate ground truth	NA	MV of 7 radiologists	NA	MV of 4 radiologists

5.4.2 Robust probabilistic framework



(a) **(b)**

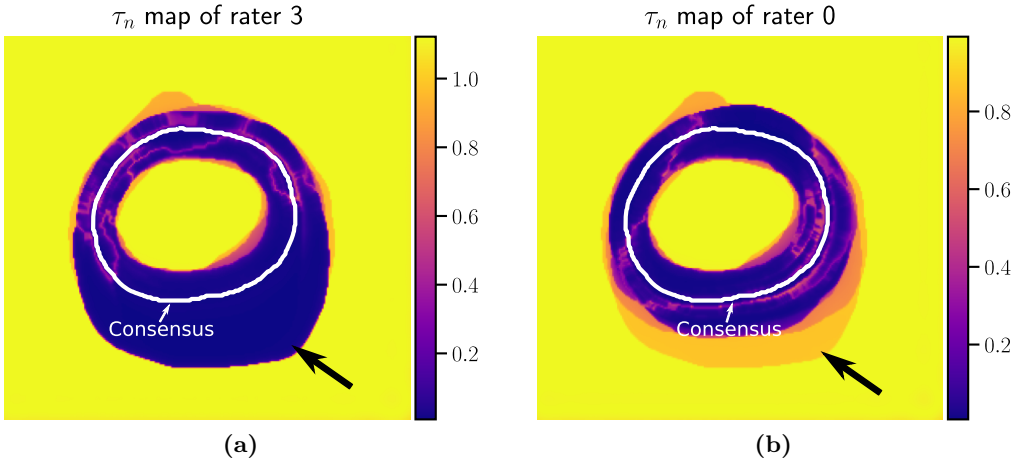
Fig. 5.5: Robust fusion of prostate binary segmentation masks using the Laplace distribution. (5.5a) Raters' manual delineations and estimated consensus shown on an axial T2 weighted image. (5.5b) Evolution of the lower bound.

Qualitative analysis.

First, we demonstrate the effectiveness of our proposed robust probabilistic model in taking into account the spatially varying performances of the raters. We consider the fusion of 7 binary prostate delineations from the ProstateBin dataset drawn by human experts into a single consensus using a framework based on the Laplace distribution fitted in 3D. The coefficient λ_s of the sigmoid function used to convert the input masks to probabilities was arbitrarily set to 5. The 7 raters segmentations and the estimated consensus are shown in Fig. 5.5a. During the inference, we maximize the lower bound, $\mathcal{L}(q)$, on the marginal log likelihood of the data. The evolution over the iterations is plotted in Fig. 5.5b.

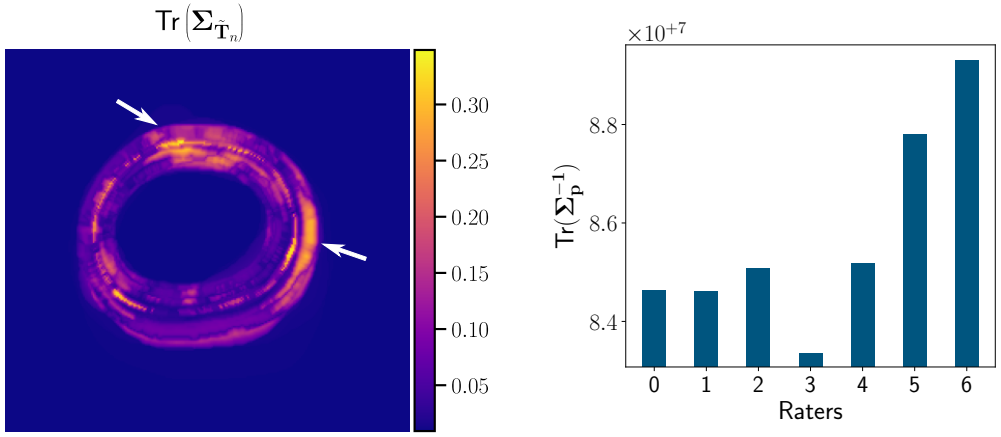
It can be seen that rater 3 seems to be an outlier with respect to the other raters at the bottom of the image, although they agree elsewhere. This local variation of the rater performance is successfully captured by the scale factor τ_n^p that spatially modulates the contribution of each rater to the consensus. In areas of poor rater performance, τ exhibits lower values which correspond to larger rater variance. Locally, raters with weak confidence will not contribute as much as others to the consensus. This is shown in Fig. 5.6a and 5.6b, where rater 3 has smaller τ_n values than rater 0 at the bottom of the image in the region highlighted by the black arrows.

The trace of the matrix $\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n}$ represents the uncertainty associated with the consensus. Low trace values correspond to a high confidence in the consensus and are typically found in area where all raters agree, as shown in Fig. 5.7a. One can observe that the highest uncertainty is not located at the bottom of the image, where there is a disagreement between rater 3 and the others, but in the image regions indicated by the white arrows. This somewhat counter-intuitive result is explained by the fact that the consensus uncertainty is estimated as a combination of the raters' precisions, weighted



(a) **(b)**

Fig. 5.6: The outlier, rater 3, exhibits locally poor performances linked to lower values of τ_n^p, in particular at the bottom of the image in the region indicated by the black arrow (5.6a). In contrast, rater 0 shows higher τ_n^p values in the same area (5.6b).



(a) **(b)**

Fig. 5.7: Uncertainty map of the consensus (5.7a). Comparison of the raters precisions (5.7b).

by the scale factor τ_n^p, as shown in Eq. 5.17. Rater 3 is considered by the model to present poorer performances in comparison to the others, which corresponds to low scale factor and precision values, as shown in Fig. 5.6a and Fig. 5.7b. Thus, rater 3 is barely taken into account for the consensus uncertainty estimation, which relies much more on the other experts. One can note that, in the Gaussian baseline model, there are no scale variables. The consensus uncertainty is a simple combination of the raters' precisions and is thus constant within the image. In particular, regions of disagreement between raters have the same level of uncertainty as regions where all raters agree. Therefore, our robust approach leads to a more realistic estimate of the consensus uncertainty, by allowing variations in the image depending on the level of agreement between raters.

The possibility of localizing visually, in a convenient manner, the most unreliable regions of the consensus is an advantage of our model in comparison to the Gaussian

baseline model, but also to the classical binary STAPLE algorithm, which does not provide any estimate of the uncertainty associated with the consensus.

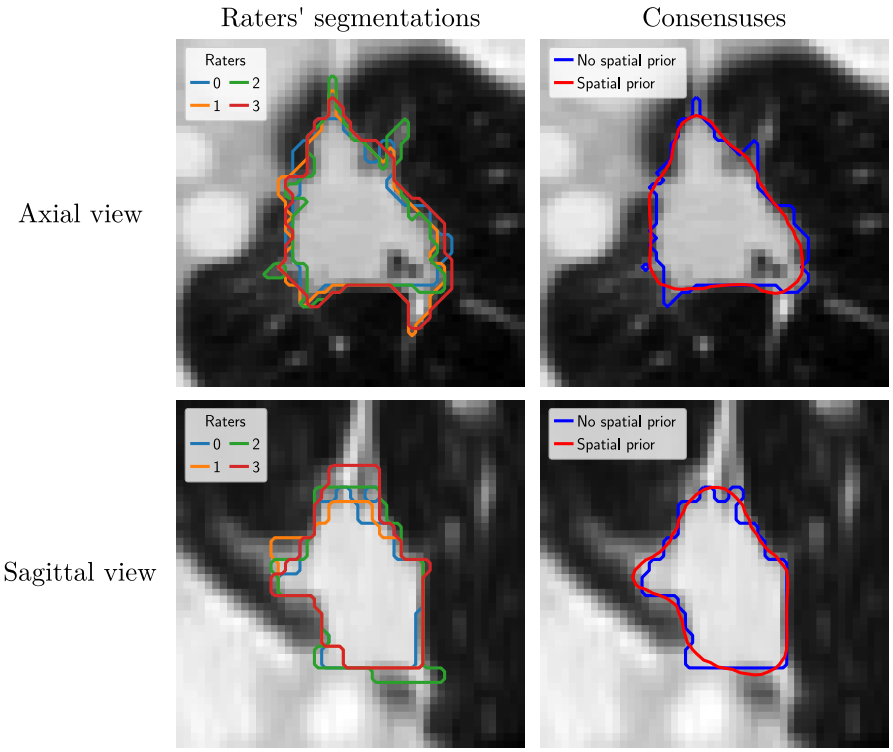


Fig. 5.8: Impact of the spatial prior on the smoothness of the consensus map obtained with the Laplace distribution on a nodule segmentation case on CT scan from the NoduleBin dataset.

One contribution of our work is the introduction of a spatial regularization prior over the consensus map. In the discrete setting, the spatial consistency of the consensus was enforced with an MRF prior, for example in [Warfield et al., 2004]. In our continuous approach, spatial correlations between voxels are taken into account by the definition of a GLSP prior over the consensus map. The key parameters are the spacing, s, between the basis function centers, the standard deviations (or radii), r, of the Gaussian functions and the position of the origin basis function. Together, they influence the level of regularization of the consensus map, large spacing and radii being associated with smoother outputs. Fig. 5.8 compares the consensuses, obtained for a nodule of the NoduleBin dataset with or without spatial regularization, in a model where the input rater maps are assumed to follow a Laplace distribution. For the model fitted with spatial regularization, the spacing, s, was set to 4 and the radius was equal to 12. The influence of the prior is clearly visible with far smoother contours.

We provide a visual comparison between the heavy-tailed distributions and the Gaussian reference in Fig. 5.9 on a nodule segmentation example from the NoduleBin dataset. The models are fitted in 3D with same spatial regularization parameters for all distributions. The inputs are the four radiologists' binary segmentations transformed to probabilities, using as before $\lambda_s = 5$ for the sigmoid function. The four manual delineations are given in the first column and the associated consensuses in the second one. It can be seen that the Student's t and GDP distributions give similar results. Both

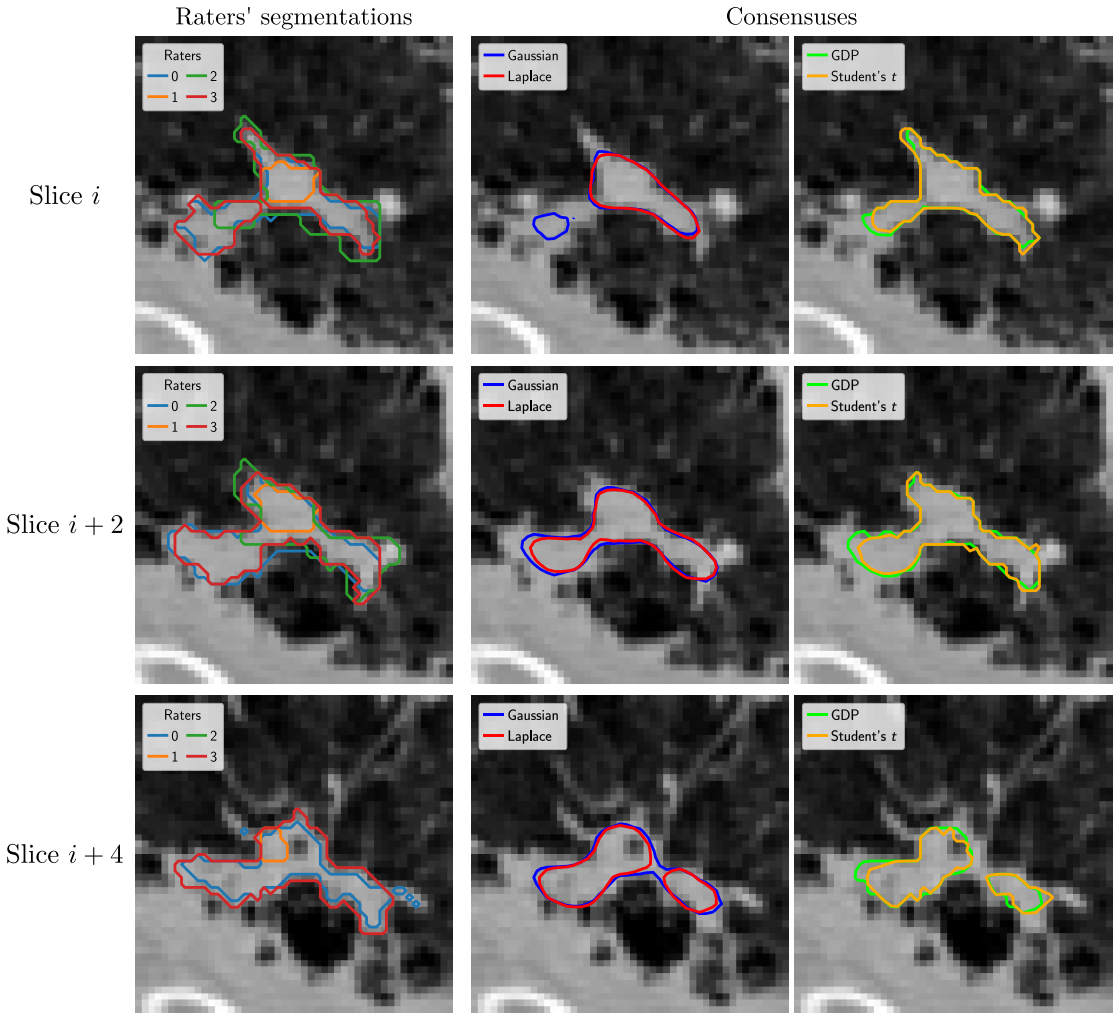


Fig. 5.9: Comparison of the heavy-tailed distributions on a nodule segmentation case extracted from the NoduleBin dataset.

have an additional degree of freedom in comparison with the Laplace and Gaussian, which allows the shape of the tail of the distribution to be adapted to the data. For this case, the mean degree of freedom ν_p between raters is 0.3 after convergence for the Student's t. It was manually set to 2 for the GDP. These values lead to heavier tails than the Laplace and Gaussian, which could explain the similar results.

The possibility of locally varying rater contributions to the consensus for the robust model leads to rater performance estimates different from those obtained with a global estimation, as for the Gaussian baseline. In Fig. 5.10, we compare the variances $(\mathbf{\Sigma}_p)_{0,0}$ corresponding to the foreground region obtained with a Gaussian or Laplace distributions. $(\mathbf{\Sigma}_p)_{0,0}$ and $(\mathbf{\Sigma}_p)_{1,1}$ can be considered the counterparts of the sensitivity and the specificity estimated in the binary setting by the STAPLE algorithm. In our framework, a large variance corresponds to poor rater performance. One can observe that the ranking between raters is close between the two distributions. However, the orders of magnitude are different with smaller variances for the robust approach. This can be explained by the fact that the experts agree in most of the image regions. The discrepancies, which contribute to poor rater performances, lie only on a small narrow band along the nodule

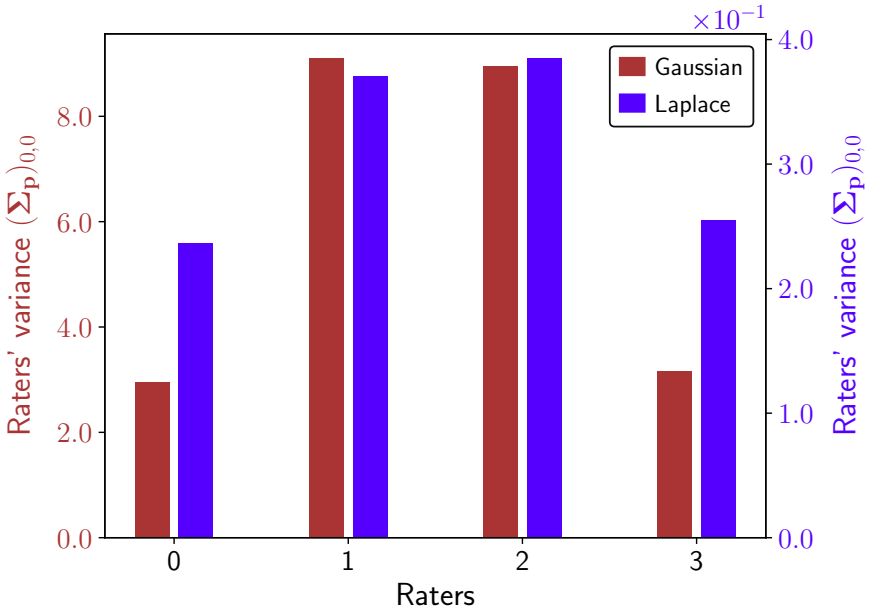


Fig. 5.10: Raters variances $(\mathbf{\Sigma}_p)_{0,0}$ corresponding to the foreground region for the models based on a Gaussian and Laplace distributions.

boundary. The local estimation of the performances allows this to be taken into account by the robust approach, which leads logically to smaller variances.

The objective of variational inference is the maximization of a lower bound over the data marginal likelihood. This lower bound can be computed to monitor the model convergence but also to provide a criteria for model selection [Blei et al., 2017]. Fig. 5.11 compares the lower bound values reached after convergence and the inference time for the different distributions on the 20 nodules from the NoduleBin dataset. The Student's t seems to lead to the highest lower bound values. This distribution has, with the GDP, an additional degree of freedom allowing the shape of the distribution to be modified and better fitted to the data. Because of numerical instabilities, this parameter is fixed manually for the GDP, whereas it is learnt automatically for the Student's t in a data-driven way. This could explain the higher lower bound values reached by the Student's t.

Regarding the computational times, the Gaussian baseline model seems to be faster, but with fewer parameters and variables to estimate. For the GDP, the expectations involving the scale factor z are evaluated with Lentz's algorithm as shown in appendix C.1, which logically leads to longer computation times.

Fig. 5.12 provides a visual comparison between our robust approach based on a Laplace distribution and models proposed in previous works. In particular, we compare our model with the original STAPLE algorithm introduced in [Warfield et al., 2002], which does not include any spatial regularization of the consensus. We also compare it to two extensions of STAPLE for continuous inputs, namely, to PSTAPLE, introduced in [Akhondi-Asl & Warfield, 2013], which uses an MRF as regularization prior, and to the continuous STAPLE algorithm, proposed in [Warfield et al., 2008], from which our approach was developed. The comparison is performed on a nodule segmentation case from the NoduleBin dataset. The inputs are therefore the delineations drawn by the

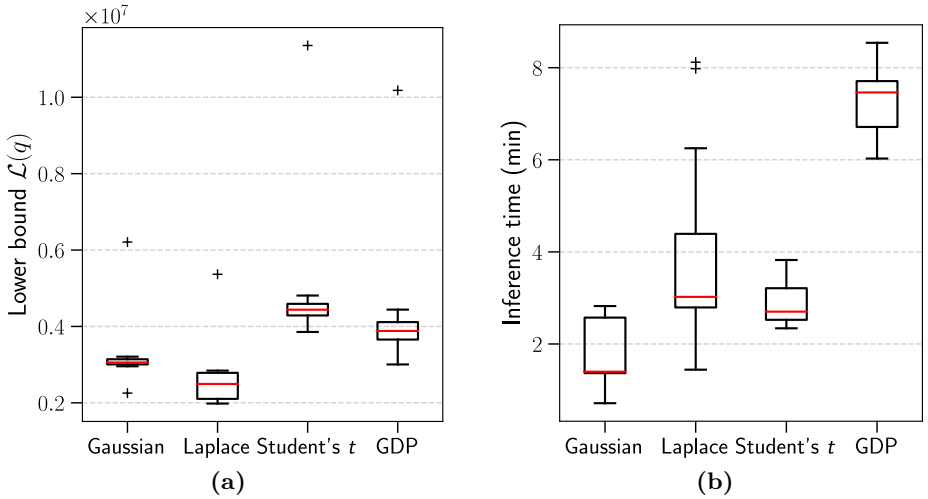


Fig. 5.11: Lower bound values reached after convergence (5.11a) and computation times obtained on the 20 nodules from NoduleBin dataset (5.11b).

radiologists and transformed to continuous maps, except for the STAPLE algorithm which handles binary inputs. For the MRF prior of PSTAPLE, a 4-connectivity neighborhood is considered, and β, the MRF hyperparameter, is set to 2. Regarding the GLSP prior used in our model, the parameters are the same as used previously.

One can observe that STAPLE and PSTAPLE lead to similar results, which could be expected as the latter is a direct extension of the former for probabilistic inputs. The effect of the MRF prior can be noted, with slightly smoother contours for PSTAPLE. The continuous STAPLE and the robust approach based on a Laplace likelihood also produce similar maps. However, our approach includes a spatial regularization prior which logically leads to smoother outputs. Moreover, the hyperparameter of the spatial prior is learnt automatically in our approach, which is an advantage in comparison with the MRF prior. Our approach is also more robust with respect to the outlier rater 1, in particular compared to the STAPLE algorithm.

Quantitative analysis.

The main difficulty when assessing the performances of data fusion algorithms is the absence of an unequivocal ground truth, which prevents any accurate quantitative comparison. This is particularly true in the medical imaging domain, where the inter-rater variability can be large.

In this section, we provide a quantitative comparison framework between our robust probabilistic approaches and methods proposed in previous works, including the most simple one, i.e., majority voting. We now consider probabilistic segmentations generated by several neural networks trained by cross-validation and tested on the NoduleNet and ProstateNet datasets, as detailed in section 5.4.1. The data fusion approaches are used to estimate a consensus between the predictions made by the different neural networks. Therefore, in contrast to the previous section, the inputs are already continuous. They need to be binarized only for the STAPLE and majority voting algorithms.

The consensuses are compared to a surrogate ground truth defined as a majority vote of the human raters' segmentations of the test set. We emphasise that, while

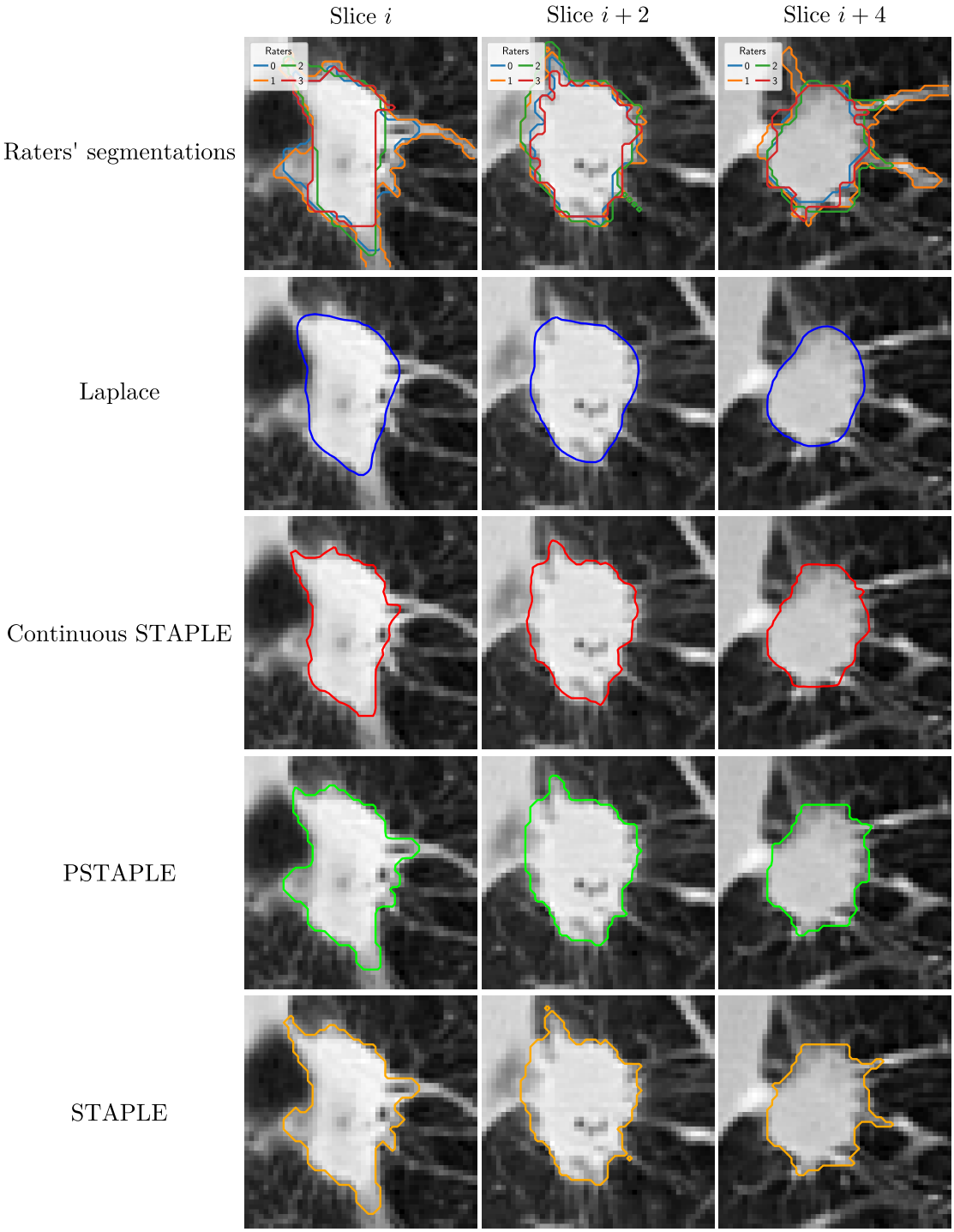
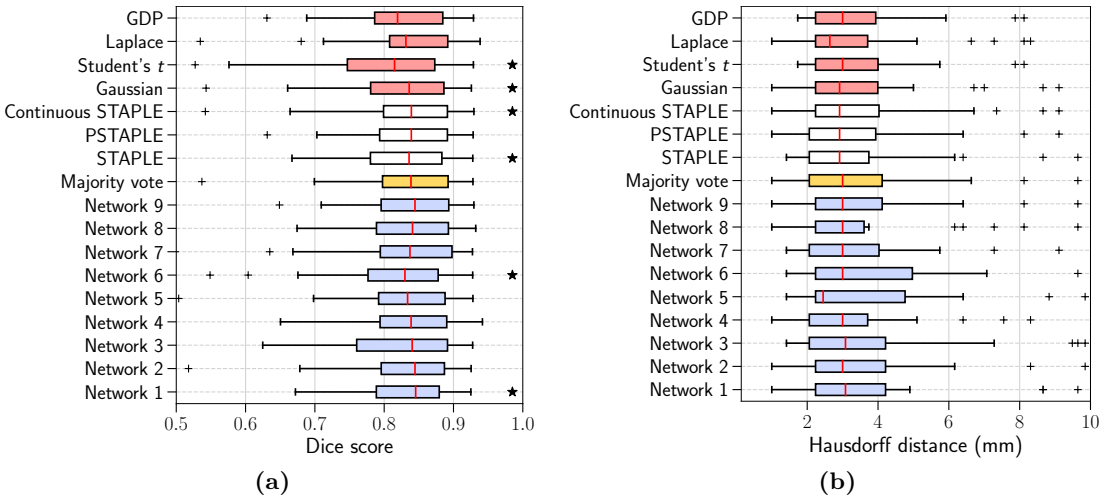
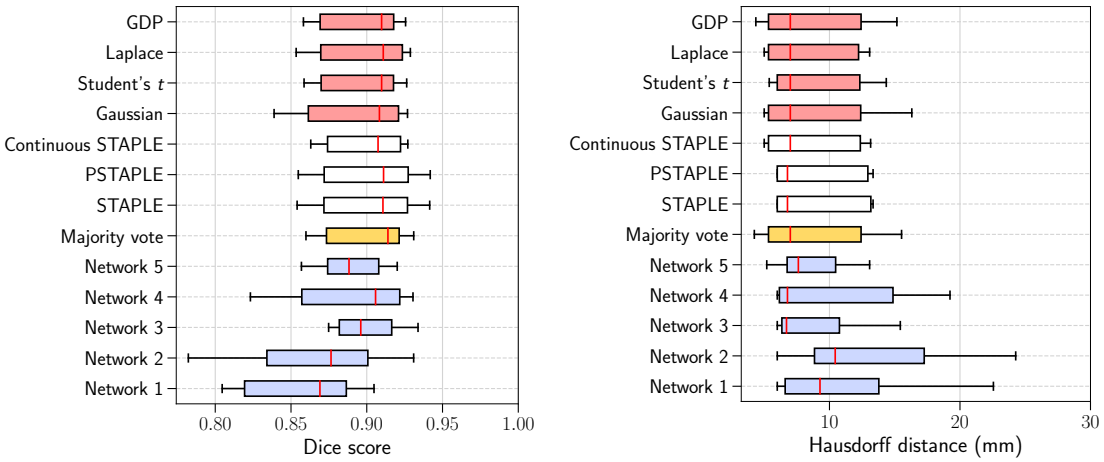


Fig. 5.12: Comparison of the robust model using a Laplace distribution with approaches proposed in previous works on a nodule segmentation case from the NoduleBin dataset.



(a) **(b)**

Fig. 5.13: Dice score (5.13a) and Hausdorff distance (5.13b) distributions over the NoduleNet dataset. Distributions marked with a $\star$ are found to be significantly different from the majority voting baseline with the Wilcoxon signed-rank test at significance level 0.05.



(a) **(b)**

Fig. 5.14: Dice score (5.14a) and Hausdorff distance (5.14b) distributions over the ProstateNet dataset.

this reference overcomes the absence of real ground truth, it is also a limitation of the comparison. We use the Dice score and the Hausdorff distance as performance metrics for the comparison. The former is a region-based metric and the latter a distance-based metric. Evaluating the performance of segmentation algorithms is a difficult task [Fenster & Chiu, 2005], and defining proper metrics remains an open challenge. Therefore, the Dice score and the Hausdorff distance may themselves be a limitation of the comparison and this should be kept in mind when analysing the results.

Regarding the lung nodules, the NoduleNet dataset contains 34 lesions, all of diameter greater than 10 mm. Smaller nodules were excluded from the analysis. The ensemble of raters is composed of 9 neural networks, whose performances on the test set in terms of

Dice score and Hausdorff distance are presented in Fig. 5.13. The Wilcoxon signed-rank test was used to compare distributions with the majority voting baseline. For the prostate dataset, 5 neural networks were used to produce probabilistic segmentations of a test set of 7 images. Thus, the prostate and nodule datasets allow us to perform a comparison between the data fusion approaches on two different structures of interest, but also with a different number of rater input maps. Fig. 5.14 shows the Dice scores and Hausdorff distances distributions for the prostate. Due to the small sample size, differences between distributions and the majority voting baseline were not tested.

First, we can observe that the neural networks of the NoduleNet dataset have more homogenous performances than those of the ProstateNet dataset. The latter were trained with a much smaller number of cases, which may explain the larger discrepancies.

Tab. 5.4: Mean Dice scores and Hausdorff distances computed between the consensuses estimated with different methods from several neural network outputs, and the reference defined as a majority vote of experts on the NoduleNet dataset.

	Dice score	Hd (mm)	Hd 95% (mm)
Majority vote	**0.83** (±0.08)	4.05 (±3.3)	2.09 (±2.29)
STAPLE	**0.83** (±0.07)	4.01 (±3.36)	2.14 (±2.3)
PSTAPLE	**0.83** (±0.07)	3.9 (±3.23)	2.07 (±2.25)
Continuous STAPLE	**0.83** (±0.08)	3.91 (±2.97)	2.19 (±2.38)
Gaussian	0.82 (±0.08)	3.86 (±2.97)	2.18 (±2.35)
Laplace	**0.83** (±0.08)	**3.51** (±2.27)	1.81 (±1.3)
Student's t	0.79 (±0.17)	3.56 (±2.08)	1.81 (±1.02)
GDP	0.81 (±0.09)	3.61 (±2.08)	**1.76** (±1.05)

Tab. 5.5: Mean Dice scores and Hausdorff distances computed between the consensuses estimated with different methods from several neural network outputs, and the reference defined as a majority vote of experts on the ProstateNet dataset.

	Dice score	Hd (mm)	Hd 95% (mm)
Majority vote	**0.9** (±0.03)	8.9 (±4.49)	4.32 (±1.6)
STAPLE	**0.9** (±0.04)	9.21 (±3.77)	4.68 (±2.27)
PSTAPLE	**0.9** (±0.04)	9.14 (±3.69)	4.66 (±2.24)
Continuous STAPLE	**0.9** (±0.03)	8.65 (±3.82)	4.26 (±1.54)
Gaussian	0.89 (±0.04)	9.11 (±4.57)	4.79 (±2.46)
Laplace	**0.9** (±0.03)	**8.6** (±3.75)	**4.22** (±1.5)
Student's t	**0.9** (±0.03)	9.06 (±3.81)	4.23 (±1.55)
GDP	**0.9** (±0.03)	8.86 (±4.39)	4.27 (±1.61)

Second, the differences between methods are small and almost never statistically significant. This is also visible in Tab. 5.4 and 5.5, which give the mean Dice score and mean Hausdorff distance for each method on the nodule and prostate datasets, respectively. In particular, the simple majority voting approach already gives good results, even better than those produced by the more complex STAPLE algorithm. Regarding our framework, better Dice scores seem to be obtained with a Gaussian distribution than with a Student's t or GDP likelihoods. In contrast, the latter two lead to smaller Hausdorff distances. The model based on a Laplace distribution appears

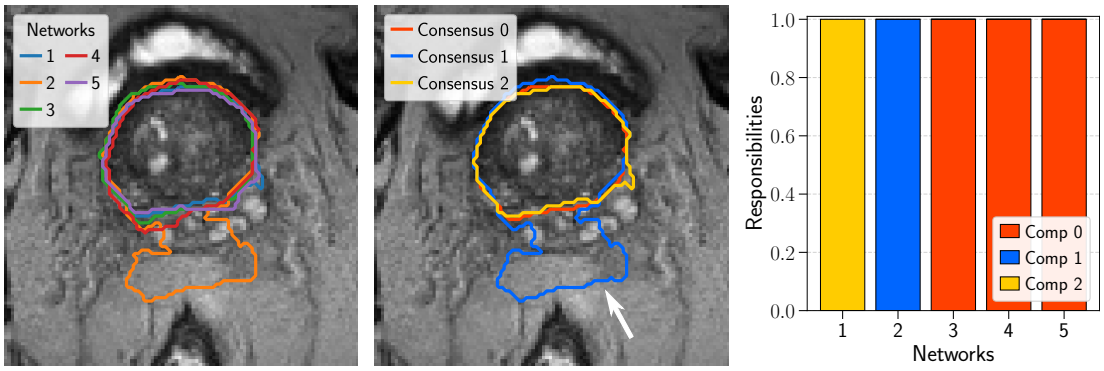
to be the most complete, as it produces balanced results between the region- and the distance-based metrics. In particular, it leads to the largest Dice scores and smallest Hausdorff distances on both datasets.

Although the differences are not statistically significant, these experiments show that our robust probabilistic framework achieves state-of-the-art results and even seems to lead to slightly better performances when the model uses a Laplace distribution.

5.4.3 Mixture of consensuses

In this last result section, we provide examples of mixtures of consensuses in Fig. 5.15 and 5.16. The inputs are the probabilistic segmentations produced by the neural networks trained by cross-validation. The mixture model is fitted on two examples extracted from the ProstateNet and NoduleNet datasets.

Fig 5.15b and 5.16b show the consensuses obtained after convergence. In both cases, three relevant contours are found. Without the mixture approach, only one consensus corresponding to the first component would have been obtained, and the regions indicated by arrows would have been ignored.

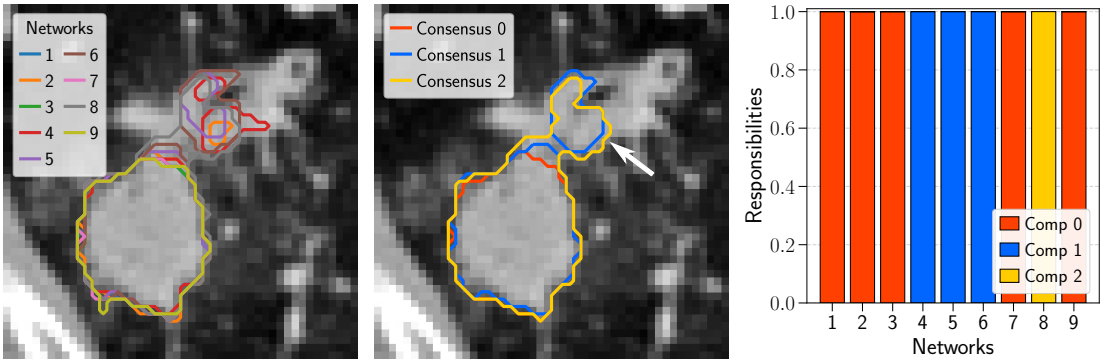


(a) **(b)** **(c)**

Fig. 5.15: Mixture of consensuses on a prostate segmentation example from the ProstateNet dataset. Input probabilistic segmentations produced by neural networks (5.15a). Estimated consensuses (5.15b). Responsibilities with 3 relevant components (5.15c).

The responsibilities are presented in Fig 5.15c and 5.16c. They indicate from which consensus each network segmentation map was generated. Thus, this method provides a novel way to cluster raters depending on their segmentations for a given image.

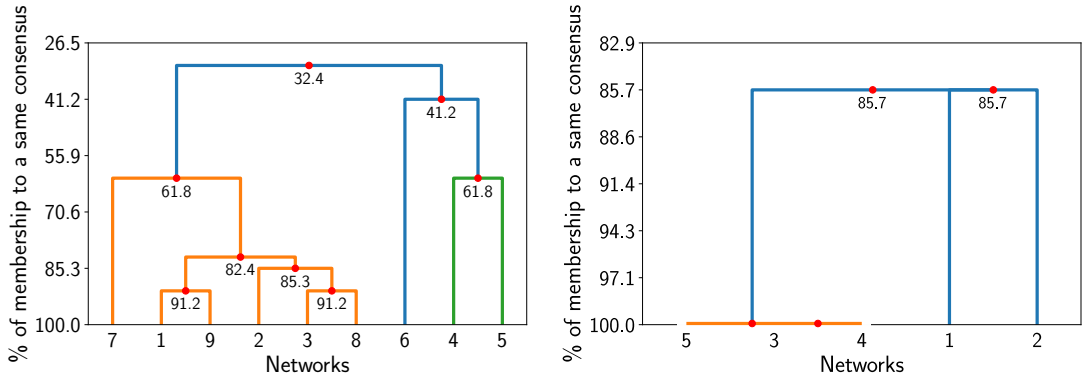
We now explore the idea of clustering raters over a batch of images, in particular over the 34 and 7 images of the NoduleNet and ProstateNet datasets. For each image of these two test sets, mixtures of consensuses were estimated and the networks were assigned to the consensus corresponding to their highest responsibility. This leads to a first clustering of the raters at the image level. Results are then aggregated over the whole test sets using hierarchical clustering with a complete-linkage approach, based on the following distance: $d(x, y) = N - N_{xy}$, where x and y denotes two raters, N is the number of segmentation cases in the dataset and N_{xy} is the number of segmentation cases where rater x and rater y are assigned to the same consensus. At each step, the two clusters having the most consensuses in common are combined. Results are presented in



(a) (b) (c)

Fig. 5.16: Mixture of consensuses for a lung nodule from the NoduleNet dataset. Input probabilistic segmentations produced by neural networks (5.16a). Estimated consensuses (5.16b). Responsibilities with 3 relevant components (5.16c).

Fig. 5.17. It shows, for example, that the network 6 is assigned to the same consensus as networks 4 and 5 in at least 41.2% of the nodule segmentation cases.



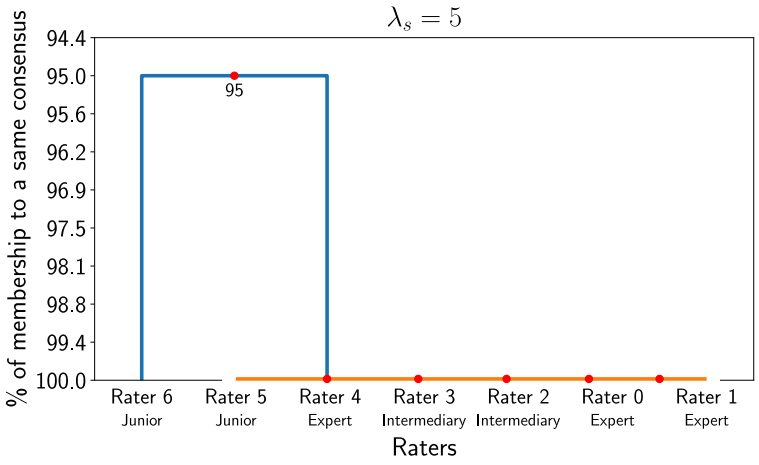
(a) (b)

Fig. 5.17: Complete-linkage clustering of the networks based on the percentage of membership of the same consensus for the NoduleNet (5.17a) and ProstateNet (5.17b) datasets.

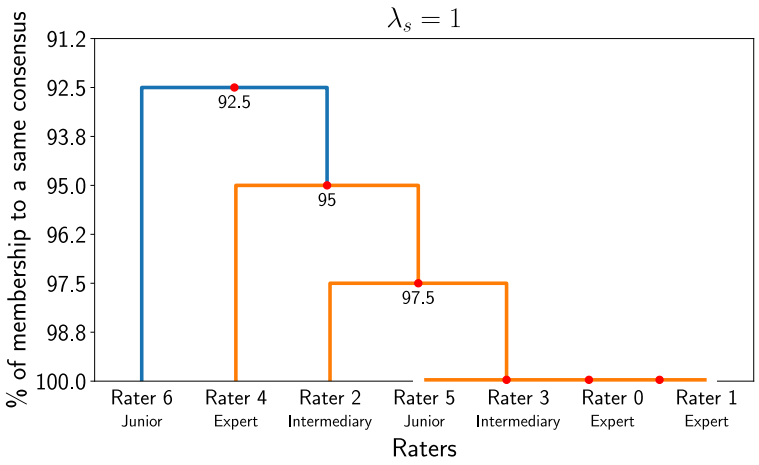
Although the networks seem to have similar performances on the NoduleNet dataset, as shown in Fig. 5.13, this approach allows two main clusters to be extracted. The group composed of networks 4, 5 and 6 appears to have a significantly different behavior than the others, as they only share 32.4% of the consensuses on the whole dataset. This difference can be visually assessed in Fig. 5.16a, where networks 4, 5 and 6 lead to a larger segmented region than the others.

The differences between networks are smaller on the ProstateNet dataset. Networks 3, 4 and 5 are always assigned to the same consensuses. In contrast, networks 1 and 2 are isolated in 14.3% of the cases. According to the results presented in Fig. 5.14, they seem to exhibit poorer performances than the others. This difference appears to be confirmed by the clustering approach.

Finally, we study the application of the mixture of consensuses for the clustering of raters for whom only binary segmentations are available. In particular, we apply



(a)



(b)

Fig. 5.18: Complete-linkage clustering of the human raters on the ProstateBin dataset, based on the percentage of membership of the same consensus, for two values of the coefficient λ_s of the sigmoid function used to convert the binary segmentation masks to continuous maps.

the approach to the segmentations drawn by the 7 radiologists on the images of the ProstateBin dataset. For each of the 40 images, a mixture of consensuses model is fitted after converting the binary segmentation masks to probabilities. We study the influence of the coefficient λ_s of the sigmoid function by presenting results for $\lambda_s = 5$, which assumes sharp transitions between image regions, and for $\lambda_s = 1$, which corresponds to a scenario with more uncertainty along the segmentation boundaries. Dendograms for the two λ_s values are shown in Fig. 5.18.

First, one can observe that the number of clusters increases with smaller λ_s values. Moreover, for $\lambda_S = 5$, 6 raters out of 7 are always grouped together over the 40 images. Rater 6 is the only one to be isolated, and only on 2 images of the dataset. This may be related to the rater lack of experience, as he/she is one of the two junior radiologists of the panel. For $\lambda_s = 1$, two other raters are extracted by the clustering approach. One has an intermediary level of experience, but the other one is considered as an expert.

This result demonstrates the applicability of our approach for studying the inter-rater variability.

5.5 Discussion

Estimating a consensus between raters is an important task in the medical imaging domain. Our work focuses on the specific problem of fusing continuous segmentation maps. It addresses three major limitations of approaches proposed in previous works, namely the estimation of the rater bias, the regularization of the consensus map and the local assessment of the rater performances.

Comparison with state-of-the-art methods showed the effectiveness of our approach. However, a limitation of the study is the definition of the ground truth used to evaluate and compare the data fusion methods. The use of majority voting in this chapter is arbitrary, and the resulting surrogate reference may actually be a flawed estimate of the real ground truth. This limitation is not specific to this chapter. It is a general problem when comparing segmentation algorithms. Yet, [Lampert et al., 2016] showed that the approach used to form the ground truth highly influences the ranking between algorithms. This problem is particularly important for medical imaging, because of the difficulty in collecting high-quality ground truths and because of the inter-rater variability. Even when the ground truth is available, for instance in the presence of numerical or physical phantoms, the metric used to assess the performances may impact the result [Fenster & Chiu, 2005; Taha & Hanbury, 2015]. How to properly evaluate the quality of segmentations remains an open issue and an interesting challenge for future work.

One contribution of our work is the introduction of a spatial prior to regularize the consensus map. In this chapter, the spatial regularity was enforced using a GLSP prior, but there are alternatives for the regularization of continuous fields. One possibility is, for example, to define a prior penalizing the total variations in the consensus map [Babacan et al., 2008]. Although specifically designed for continuous inputs, our data fusion approach can handle binary segmentations, once they are converted to probabilities. In this chapter, we used a transformation based on a Euclidean distance map and the sigmoid function. Varying the value of the parameter λ_s of the sigmoid leads to different consensus estimates by allowing various levels of uncertainty to be simulated. It is an advantage in comparison to the discrete data fusion methods, which neglect uncertainty by always assuming sharp transitions between image regions. However, one limitation of the approach followed in this chapter is that the coefficient is independent of the location, leading to equal levels of uncertainty along the segmentation boundaries in all image regions. It could be improved by varying the slope of the transitions depending on the location and contrast in the image, allowing, for example, more uncertainty to be assumed in areas where raters disagree.

Our approach provides a statistical framework for assessing the performances of the raters. In particular, the mixture of consensuses model is a novel approach to study the inter-rater variability, cluster raters and detect outliers. The approach, inspired by variational boosting, allows the appropriate number of consensuses to be estimated in a data-driven way. It requires a reduction of dimension, performed in this chapter by PCA. This method maximizes the variance of the data projected in the latent space, which is an attractive property when the objective is to identify patterns among raters. However,

other reduction techniques could be used, and their investigation represents an avenue for future work. Moreover, the mixture of consensuses was only tested with a simplified model assuming a Gaussian likelihood and no rater bias. This model could be extended by adding a bias for the raters and replacing the Gaussian with a robust distribution. However, in contrast to the classical robust model, the rater bias would not be directly related to over- or under-segmentation anymore, because of the projection into the latent space. Similarly, it would not be possible to connect in a straightforward manner the variations in the scale factor to specific locations in the image, making the model less interpretable. Furthermore, the mixture of consensuses with a Gaussian distribution is already more robust than the Gaussian model with a unique component, in particular by allowing outliers to be isolated. We note that a related approach for outlier detection was proposed by [Commowick & Warfield, 2009]. However, it is purely based on a statistical comparison of the raters' biases and variances and does not allow several consensuses to be generated.

Another interesting topic of research is the evaluation of the intra-rater variability, which reflects the consistency of a rater when segmenting the same image several times. This could be assessed using our model, by fusing the different segmentations of an image produced by a rater and sharing the variance $\mathbf{\Sigma}_p$ between the input maps. After convergence, this parameter would give an estimate of the intra-rater variability.

Moreover, the experiments in this chapter were designed such that the raters performances were evaluated independently from one image to another. Yet, it is reasonable to assume that part of a rater's performance does not depend on a given image. For example, errors related to a lack of experience may be repeated over a whole set of images. In order to take this observation into account, one possibility would be to add a prior over the rater performance parameters, and then learn the prior hyperparameters using several segmentation cases. One can assume that this strategy would lead to a more robust estimate of the raters' performances.

This approach could also be followed to constrain the scale factor to take more uniform values between the raters. Indeed, we can see on Fig. 5.6 that, although raters 0 and 3 agree in the corner of the image, their τ_n^p values are not equal. This does not mean that they do not contribute equally to the consensus in these image regions, as each rater contribution also depends on the rater variance. However, more uniform scale factor values could be obtained by the introduction of a prior and sharing its parameters between the raters.

5.6 Conclusion

Consensus estimation between raters is an important but difficult problem. The main challenge is to assess the performance of each rater and the associated uncertainty properly. Many approaches have been proposed to address this challenge for discrete inputs. In contrast, the continuous setting has received less attention.

In this chapter, we focused on this latter case and proposed a novel robust Bayesian framework for the fusion of continuous segmentation maps based on heavy-tailed distributions. A major contribution of our work is the local assessment of the raters performances, which were only estimated globally in previous approaches. These locally varying performances are made possible by the writing the heavy-tailed distributions as Gaussian scale mixtures. Moreover, the spatial consistency of the consensus is enforced by

the introduction of a regularization prior. We propose a convenient inference framework based on variational calculus that allows the model variables and parameters to be estimated in a data-driven way.

Consensuses obtained with the heavy-tailed distributions were visually compared and this qualitative comparison demonstrated that the distributions lead to different segmentation results. A quantitative comparison with methods proposed in previous works was performed using probabilistic segmentations generated by neural networks. We showed that our approaches achieved state-of-the-art results. In particular, the model fitted with a Laplace distribution led to slightly better performances, both for the region- and distance-based metrics.

This chapter also explores the novel concept of mixtures of consensuses. Unlike classical approaches, several consensuses can be obtained, which highlight the potential presence of several patterns among raters. This model also provides a novel way to cluster raters, allowing outliers to be extracted.

Several ideas to extend our framework were developed in the discussion. In particular, applying our framework to several segmentations generated by a rater on the same image to study the intra-rater variability seems to be a promising research avenue for future work.

In conclusion, we believe our method may be a useful tool to estimate a consensus between several segmentation maps, and the approach could be of interest in other fields of application where data fusion is required.

Acknowledgments

This work was partially funded by the French government, through the UCA$^{\text{JEDI}}$ and 3IA Côte d'Azur "Investments in the Future" projects managed by the National Research Agency (ANR) with the reference numbers ANR-15-IDEX-01 and ANR-19-P3IA-0002 and supported by the Inria Sophia Antipolis - Méditerranée "NEF" computation cluster. Data were partially extracted from the Clinical Data Warehouse of the Greater Paris University Hospitals (Assistance Publique – Hôpitaux de Paris).

Chapter 6

End-to-end analysis of a computerized lung cancer screening pipeline based on LDCT

Contents

6.1 Introduction . . . 106
6.2 Material . . . 109
6.2.1 Training and validation: the LIDC-IDRI dataset . . . 109
6.2.2 Independent test sets . . . 112
6.3 Lung cancer screening pipeline . . . 113
6.3.1 Framework overview . . . 113
6.3.2 Lung segmentation network . . . 115
6.3.3 Nodule detection network . . . 116
6.3.4 Characterization network . . . 119
6.4 Results . . . 121
6.4.1 Cross-validation results on the LIDC-IDRI dataset . . . 121
6.4.2 Tests on independent datasets . . . 126
6.5 Discussion . . . 128
6.6 Conclusion . . . 132

Lung cancer is the leading cause of death by cancer. Large scale studies have shown the potential of low-dose computed tomography (LDCT) screening with a reduction in lung cancer mortality. Yet, lung nodule detection is one of the most tedious and time-consuming task for radiologists. There is therefore a need in developing automated tools, not to replace but to assist the radiologists, in order to enable the implementation of lung cancer screening policies at large scales. An automated lung screening pipeline is usually composed of a few steps, starting with the nodule detection followed by a false positive reduction step, the characterisation of the detected candidates and finally a prediction at the scan level. The development of methods corresponding to each of these steps has been fostered by the public release of the LIDC-IDRI database, which led to major advances in the field of computer-aided lung cancer screening. However, most of prior works are dedicated to a particular task, focusing on the specific challenges associated with it, but few present results after integrating the whole pipeline. The first objective of this chapter is to provide a comprehensive analysis of a computerized

lung cancer screening pipeline, from the detection and characterization of nodules to the final prediction at the patient level. Moreover, most of prior results were obtained on the LIDC-IDRI database, which became a popular benchmark for the comparison of algorithms. Yet, this LIDC-IDRI dataset has itself some limitations due to the annotation process, which raises the question of the true performances on real life clinical data. The second objective of this chapter is to investigate this generalization issue by applying our pipeline trained solely on the LIDC-IDRI data on three independent test sets, i.e., on a subset derived from the NLST database, on the training data of the 2017 Data Science Bowl, and on a private database of COPD patients.

Part of this chapter corresponds to the following publication:

- [Audelan et al., 2021] **B. Audelan**, S. Lopez, P. Fillard, Y. Diascorn, B. Padovani and H. Delingette. Validation of lung nodule detection a year before diagnosis in NLST dataset based on a deep learning system. *Submitted to ERS International Congress 2021.*

6.1 Introduction

In 2020, lung cancer was responsible for 135720 deaths in the United States, far ahead from the other types of cancer for both men and women [Siegel et al., 2020]. In France, there was 33117 deaths by lung cancer in 2018 [Defossez et al., 2019]. It is a major public health problem in the world with economical consequences. If smoking remains the main risk factor for lung cancer, other factors such as exposition to air pollution are now commonly admitted [Barta et al., 2019]. Patient prognosis heavily depends on the time of detection, as early stage diagnosis improves dramatically the chances of survival [Torre et al., 2016].

Two large randomized controlled studies, the American National Lung Screening Trial (NLST) and the Dutch-Belgian NELSON trial, revealed the effectiveness of lung cancer screening with low-dose computed tomography (LDCT), with a positive impact on overall survival [NLST, 2011; Koning et al., 2020]. These results were confirmed by the 39% reduction in lung cancer mortality observed more recently in the Multicentric Italian Lung Detection (MILD) trial [Pastorino et al., 2019]. Lung cancer screening in LDCT scans involves to detect and characterize lung nodules, i.e. to evaluate their malignancy. Lung nodules are small, approximately round, lesions of diameter less than 3 cm [Ost et al., 2003]. They are common findings in chest CT scans in practice, and if some of them may be suspicious for lung cancer, the majority are benign. For instance, an analysis of the results of 8 large screening trials performed by [Wahidi et al., 2007] showed that up to 51% of the participants presented nodules, but that the prevalence of lung cancer in subjects with nodules was at most 12%. Therefore, the small size of the lesions in comparison with the scan volume and the high rate of false positives make the detection and characterization of nodules one of the most challenging radiological task [Rubin, 2015]. Moreover, this task is associated to a high inter-rater variability and a sensibility dropping significantly for small diameter nodules [Rubin, 2015].

LDCT is already the cornerstone of lung cancer screening in the United States [Wood et al., 2018]. It follows that, with the implementation of LDCT screening at large scales, there is a need to develop computer-aided methods in order to assist clinicians. The LIDC-IDRI database was released in order to encourage the development of automatic tools for lung cancer detection [Armato III et al., 2011]. More than a

thousand computed tomography (CT) scans were made publicly available, along with annotations provided by radiologists. This release led to major advances and resulted in numerous publications [Pehrson et al., 2019]. In particular, it allowed the LUNA16 challenge to be organized, which used the LIDC-IDRI data to provide a comparison framework for detection algorithms [Setio et al., 2017]. The value of the LIDC-IDRI dataset lies in the detailed annotations provided by 4 radiologists that include the spatial coordinates of the lesions, a characterization of their appearance and an assessment of their malignancy. The multiplicity of experts allows moreover the inter-rater variability to be estimated. However, the main limitation of these annotations is that they are purely based on radiological criteria and were not confirmed by any histopathological analysis. As highlighted above, the radiological evaluation of nodules is a very difficult task, prone to the reader's subjectivity. This is particularly the case for the malignancy assessment because of significant overlaps in features characterizing benign and malignant lesions [Erasmus et al., 2000; Snoeckx et al., 2018]. This limitation is moreover exacerbated by the absence of a ground truth cancer status for the subjects [Armato III et al., 2011].

A computer-aided lung screening pipeline is usually made of several steps. It starts with the pulmonary nodule detection, which involves to examine the CT scan looking for abnormalities. Most lesions not being cancerous, the challenge is to achieve a high level of sensitivity while keeping the false positive rate at a reasonable level [Ost et al., 2003]. The second challenge is the small size of the structures to be detected: with a diameter up to 30 mm but mostly lower than 10 mm, they usually represent less than 0.013% of the image volume, making their detection subtle [Gould et al., 2007; Rubin, 2015]. A wide range of approaches has been proposed for automatic lung nodule detection, including feature-based algorithms [Naqi et al., 2018; Bai et al., 2015; Shaukat et al., 2017] or deep learning-based methods [Jaeger et al., 2018; Winkels & Cohen, 2019; Xie et al., 2019a]. For a detailed survey of nodule detection methods applied to the LIDC-IDRI database, see [Pehrson et al., 2019].

Among the candidates extracted by the nodule detection step, many are generally false positives. An efficient lung cancer screening pipeline is expected to achieve a high level of sensitivity to retrieve all cancer cases, while not neglecting the specificity in order to avoid any unnecessary anxiety related to false positive candidates. Therefore, the detection is usually followed by a false positive reduction step to filter out the wrong candidate locations. The difficulty now lies in the wide variety of nodules, in terms of appearance (ground-glass, part-solid or solid classification) but also of morphological characteristics (size, shape, margin), which are highly variable [Gould et al., 2007]. In addition, there are many lung structures resembling nodules making the differentiation challenging. Several works focusing on the false positive reduction task have been published, proposing feature or deep learning-based approaches [Ge et al., 2005; Setio et al., 2016; da Silva et al., 2018].

The final step of the pipeline is the classification of the selected candidates depending on their suspected malignancy. The malignancy assessment of a nodule is a challenging task and only biopsy can provide a definitive diagnosis. However, some characteristics like spiculated margins, wall thickness or large diameters are rather associated with malignant patterns whereas others like calcification are indicators of benign lesions [Ost et al., 2003]. Classification frameworks were for instance proposed in [Xie et al., 2019b; Xie et al., 2018; Wu et al., 2018; Causey et al., 2018].

Last but not least, the final step is the assembling of the whole pipeline, which starts with a CT scan as input and outputs the suspicious nodules, if any. The aggregation of

the results at the nodule level allows then finally to produce a decision at the patient level. In the automated lung cancer screening literature, most works, including for instance those cited above, do not implement all steps. Instead, they focus on a particular task and the challenges associated with it. However, to be able to consider real-life applications, a proper end-to-end assessment of the framework is necessary, as the performance of each step also depends strongly on the preceding results. In contrast to the numerous papers addressing a specific task, few are analysing the impact of combining all of them, in particular on the LIDC-IDRI dataset, as noted by [Bonavita et al., 2020].

One objective of this chapter is to fill this gap by providing a comprehensive analysis of an end-to-end lung screening pipeline. [Zhu et al., 2018; Bonavita et al., 2020; Ozdemir et al., 2020; Ardila et al., 2019; Zhang et al., 2019; Liao et al., 2019] are related works that also propose complete pipelines and are discussed hereinafter.

In [Zhu et al., 2018], the authors propose a framework trained and evaluated on the LIDC-IDRI dataset. They use the lung segmentations provided by the LUNA16 challenge [Setio et al., 2017], thus ignoring potential problems due to the lung segmentation step in the final analysis. Moreover, results are solely given in terms of accuracy. Yet, accuracy is not an appropriate measure of performance for extremely imbalanced datasets [He & Garcia, 2009], which is the case here with a minority of true positive cancer cases.

[Bonavita et al., 2020] introduces another pipeline also trained and tested exclusively on the LIDC-IDRI data. The decision at the scan level involves training another classifier in addition to the nodule characterization model, making the whole framework more complex and the final decision less interpretable. In contrast, in [Zhu et al., 2018], a scan was automatically considered as a cancer case as soon as one nodule was identified as suspicious. Moreover, an important limitation shared by [Zhu et al., 2018] and [Bonavita et al., 2020] is the absence of any independent test set with unequivocal annotations. As mentioned above, the malignancy scores provided in the LIDC-IDRI database are only based on a radiological assessment that was not confirmed by biopsy, and the cancer status of the subjects is not available.

In [Liao et al., 2019; Ozdemir et al., 2020], pipelines are evaluated on the stage 2 data of the Data Science Bowl (DSB) 2017 competition [DSB, 2017], after being trained on a combination of the LIDC-IDRI data and of the DSB stage 1 data. Thus, even if the DSB images originate from multiple sources, the evaluation is not performed on a completly independent test set. Moreover, the DSB data provides a label at the scan level but none at the nodule level, which prevents any complete comparison.

Finally, [Zhang et al., 2019] pre-trained their framework using the LIDC-IDRI and the DSB data and performed the evaluation on an independent proprietary dataset. [Ardila et al., 2019] trained and tested their method on data from the NLST and on an independent test set. However, both of them do not provide a detailed result analysis of the nodule detection and characterization steps.

In this chapter, we conduct a comprehensive analysis of a lung cancer screening pipeline. Using the widely used data splitting approach proposed by the LUNA16 challenge, we present cross-validation results on the LIDC-IDRI dataset, up to the patient level which was, to the best of our knowledge, never proposed before. As pointed out previously, the radiological nature of the annotations are a limitation of the LIDC-IDRI data when it comes to the nodule characterization and the scan label. Therefore, we also evaluate our pipeline trained on the full LIDC-IDRI dataset on three independent test sets: the DSB stage 1 data containing images of 1595 subjects, a subset of the NLST database corresponding to 1179 patients [NLST, 2011] and a private database of 610

patients with chronic obstructive pulmonary disease (COPD), denoted as AIR [Leroy et al., 2017].

In contrast to [Liao et al., 2019; Ozdemir et al., 2020], we use the DSB stage 1 data only for testing. The availability of the cancer status enables an evaluation of the pipeline prediction at the scan level. The second test set is a subset of 1179 patients from the NLST study, for which 2 CT scans are available one year apart. The diagnosis for the cancer cases was established in the last year, thus enabling to apply the pipeline both on the image at diagnosis and on the image one year before, and therefore to evaluate the algorithm ability to make predictions one year before the radiologists. In addition to the subject cancer status, the NLST data provides nodule level annotations which include the localization of the lesions and their malignancy classification confirmed by biopsy. This allows us to study the generalization capability of a framework trained only on radiological annotations to real life screening data. The third test set is also the most recent. It is a collection of LDCT scans of subjects with COPD, a chronic disease corresponding to an inflammation of the bronchii. It leads to a progressive obstruction of the airways and a deterioration of the lungs, causing difficulties in breathing. It has been reported as a potential increased risk factor of lung cancer [Durham & Adcock, 2015]. This test set enables to evaluate the robustness of the pipeline to more recent images and to study the impact of lung comorbidities on the detection results.

Our pipeline, entirely based on deep learning, is composed of three steps, each one associated with a specific network. The number of steps and models involved is smaller than what was proposed in some previous works, for instance in [Bonavita et al., 2020; Ardila et al., 2019]. It allows the framework to be trained more easily and also more interpretable.

The main contributions of this chapter are summarized below:

- First, we provide a comprehensive analysis of an end-to-end lung screening pipeline on the LIDC-IDRI dataset using cross-validation.
- Second, we apply our framework trained on the LIDC-IDRI dataset on three independent test sets, including NLST data. It enables us to study the performance of a pipeline trained solely on radiological annotations on data with biopsy-confirmed labels. We also present results on images one year before the clinical diagnosis, and test the robustness of the pipeline to images of patients with lung comorbidities.

The rest of the chapter is organized as follows. In section 6.2, we introduce the datasets used for training, validating and testing our lung screening pipeline. Section 6.3 begins with an overview of the whole lung screening pipeline and then provides a more detailed description of each step. Finally, section 6.4 gathers the experiments and results.

6.2 Material

6.2.1 Training and validation: the LIDC-IDRI dataset

Our proposed lung screening pipeline was trained and validated using the publicly available LIDC-IDRI database. The dataset is a retrospective collection of 1018 LDCT scans with 0.6 to 5.0 mm slice thickness, coming from a collaboration between 7 medical centers and 8 medical imaging companies. The in-plane pixel size ranges from 0.5 to 1.0

mm. Inclusion criteria were based on the image quality and are detailed in [Armato III et al., 2004].

Each scan in the LIDC-IDRI database was reviewed by 4 radiologists in a 2 stage annotation procedure described in [McNitt-Gray et al., 2007]. Annotations for lesions of sizes ranging from 3 mm to 30 mm include localization, manual delineation and characterization with respect to several properties, in particular malignancy and texture, each noted between 1 and 5. All annotations are solely based on radiological criteria and do not include any histological analysis.

Among the annotations, the malignancy score s_m aims to estimate the malignancy likelihood of the nodule between 1 (highly unlikely) and 5 (highly suspicious). It is a subjective assessment from the radiologists and was not confirmed by biopsy. The texture score s_t classifies the nodule appearance from ground-glass to solid. Examples of nodules exhibiting various texture and malignancy scores are presented in Fig. 6.1 and 6.2.

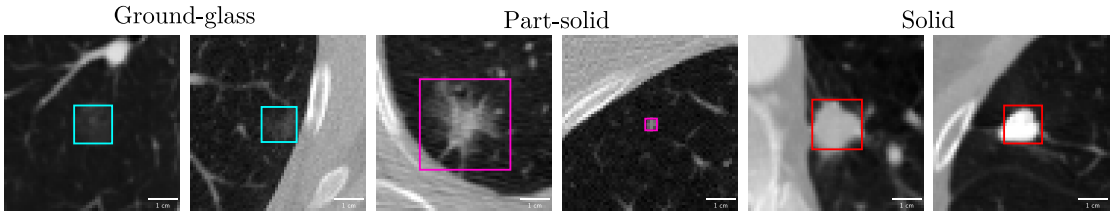


Fig. 6.1: Texture characterization of nodules: examples of ground-glass, part-solid and solid nodules.

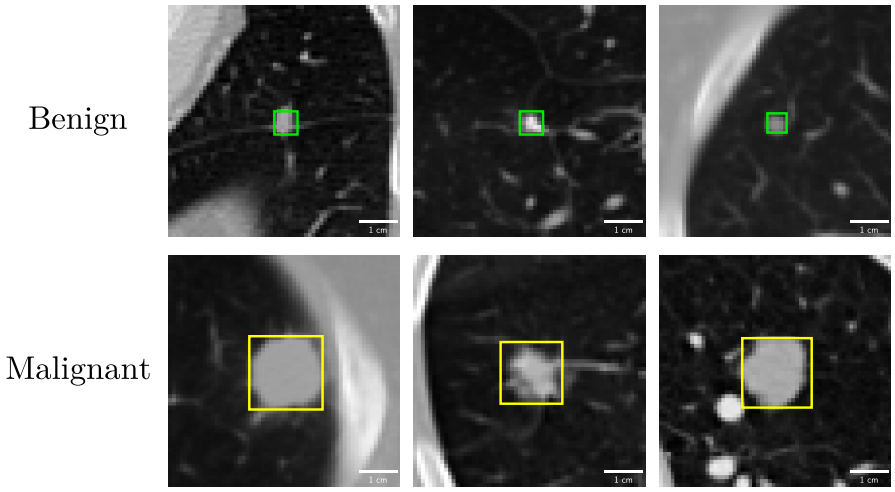


Fig. 6.2: Malignancy characterization of nodules: examples of benign and suspicious nodules.

In this chapter, we worked with a refined version of the LIDC-IDRI data defined in the LUNA16 challenge dedicated to the automatic detection of pulmonary nodules [Setio et al., 2017]. Scans with a slice thickness larger than 3 mm, with inconsistent spacing or missing slices, were excluded by the organizers, leading to a total of 888 CT scans. Nodule annotations were clustered using the *pylidc* API [Hancock & Magnan, 2016] giving a total of 2281 nodules.

To deal with the inter-rater variability, feature consensuses between the 4 radiologists were obtained by averaging the scores and rounding to the nearest integer. Moreover, we followed the reference standard of the LUNA16 challenge and kept only nodule annotations accepted by at least 3 out of 4 radiologists, which reduced the total number of nodules to 1186.

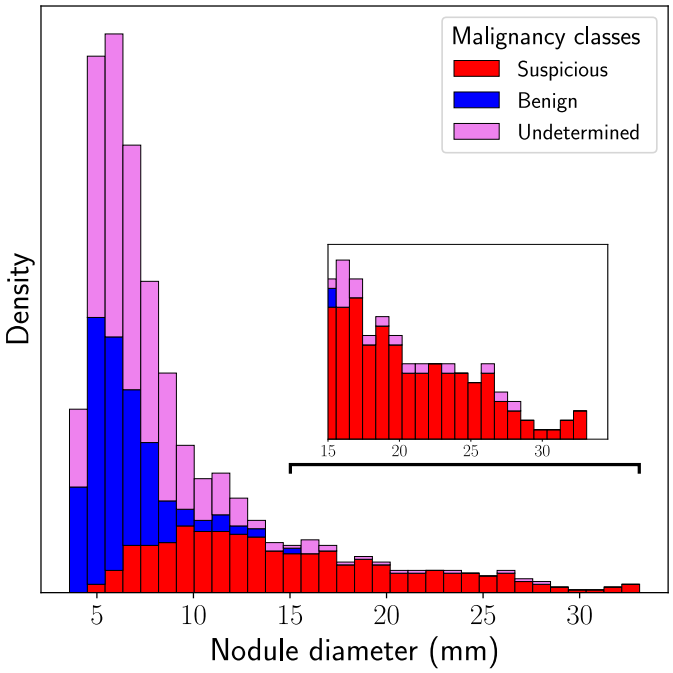


Fig. 6.3: Normalized histograms of the nodule diameter in the LIDC-IDRI dataset for the 3 malignancy classes.

Regarding the malignancy assessment, nodules with an intermediate malignancy score, $2 < s_m < 4$, are considered as ambiguous (undetermined malignancy) and were not taken into account when evaluating the cancer predictions of the pipeline. Others are classified into benign or suspicious depending on their score, if $s_m \leq 2$ or $s_m \geq 4$, respectively. The distribution of the nodule diameters for the 3 malignancy classes (suspicious, benign and undetermined) is shown in Fig. 6.3. Large diameters are associated with an increased likelihood of malignancy. Nodules within the range 2 to 15 mm are the most challenging to evaluate with an overlap of the 3 classes.

Cancer diagnoses at the patient level are not provided in the LIDC-IDRI database. In order to perform a complete analysis of our pipeline, we propagated labels from the nodule level to the scan level according to the following rule: a scan is labelled as cancer if at least one of its associated nodules was identified as suspicious by the radiologists. This rule, used to establish the subject cancer status, together with the absence of biospy-confirmed labels for nodules, are limitations of the study on the LIDC-IDRI dataset. A summary of the LIDC-IDRI data used in this chapter is presented in Fig. 6.4.

Our lung screening pipeline was trained and validated on the LIDC-IDRI dataset following the widely used 10-fold cross-validation scheme proposed by the LUNA16 challenge. This scheme splits the data into ten subsets of equal size on a patient level, and enables comparison with existing results.

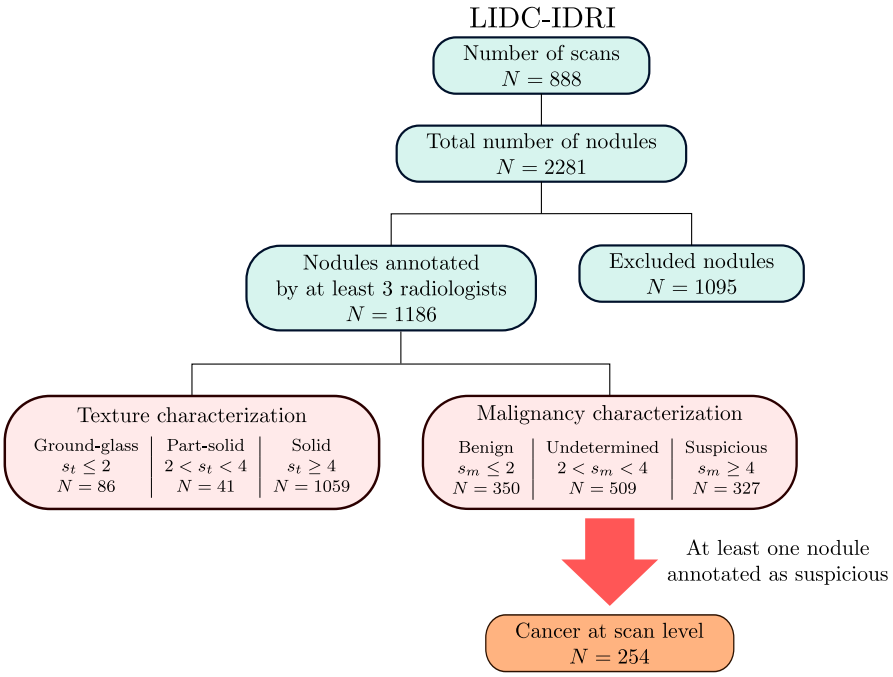


Fig. 6.4: Description of the LIDC-IDRI dataset used to train and validate the lung screening pipeline.

6.2.2 Independent test sets

The NLST subset.

The American National Lung Screening Trial (NLST) was a large randomized controlled study whose objective was to determine the effectiveness of computed tomography for lung cancer screening in comparison with chest radiography. 53454 participants were enrolled in the United States between 2002 and 2004 according to the following eligibility criteria: being aged between 55 and 74 years old, with a smoking history of at least 30 pack-years, and, if former smoker, having quitted smoking within the last 15 years.

26722 subjects were randomly assigned to the LDCT arm of the study, and underwent three screening tests at one year of interval, denoted as T_0, T_1 and T_2, between 2002 and 2007.

In this chapter, we consider a subset of the NLST dataset composed of 1179 patients from the LDCT arm of the trial, including 177 lung cancer cases, all diagnosed at T_1 or T_2. A scan one year before diagnosis is therefore available for all cancer patients. In this subset, the slice thickness varies between 0.02 to 5.0 mm. The in-plane pixel size ranges from 0.5 to 0.9 mm.

This subset is used to perform two experiments. First, we evaluate the performance of the pipeline on the T_2 image of the 1002 cancer-free subjects, and on the image at diagnosis (T_1 or T_2) for the 177 cancer patients. 2352 nodules were identified by the radiologists on this set of 1179 images, including 177 lesions confirmed to be malignant after biopsy. In a second step, the pipeline is applied on the image one year before diagnosis for the cancer patients, in order to evaluate its ability to detect cancer one year before radiologists.

Tab. 6.1: Summary of the main characteristics of the 4 databases used in this chapter.

	LIDC-IDRI	NLST	DSB stage 1	AIR
Used for	Training and validation	Testing	Testing	Testing
Acquisition date	NA	2002 - 2007	NA	2015 - 2018
# of subjects	888	1179	1595	610
# of cancer patients	254	177	495	22
# of nodules	1186	2352	NA	NA
# of malignant nodules	327[1]	177[2]	NA	NA

[1] Based on a subjective radiological assessment.
[2] Confirmed by biopsy.

The DSB stage 1 data.

The second test set is the training data from the 2017 Data Science Bowl (DSB), denoted as stage 1. It contains 1595 CT scans of high-risk patients, coming from multiple sources, including the NLST database. The slice thickness and the in-plane pixel size range from 0.6 to 2.5 mm and from 0.5 to 1.0 mm, respectively. Among the 1595 subjects, 495 are labelled as cancer. The subject cancer status is the only annotation provided in this dataset.

The AIR cohort.

The last test set is a private database denoted as AIR. The AIR project was a prospective study conducted by 21 medical centers in France between 2015 and 2018, with the objective of assessing the role of chest CT and circulating tumor cells in lung cancer screening. Eligible participants were above 55 years old, had a smoking history and suffered from chronic obstructive pulmonary disease (COPD).

Participants were invited to undergo 3 screenings at one year of interval. However, we consider in this chapter only the scans collected during the first screening test, where 22 subjects were diagnosed with lung cancer. The slice thickness and the in-plane pixel size range from 0.3 to 4 mm and from 0.3 to 1.0 mm, respectively. As for the DSB stage 1 data, the subject cancer status is the only available annotation.

The main characteristics of the 4 datasets used in this chapter are reported in Tab. 6.1. One can observe that the combined test datasets are larger than the LIDC-IDRI training set and correspond to distinct populations. A graphical representation of our training and testing framework is proposed in Fig. 6.5.

6.3 Lung cancer screening pipeline

6.3.1 Framework overview

The proposed lung screening pipeline is composed of the following steps:

1. Image pre-processing,
2. Lung segmentation,

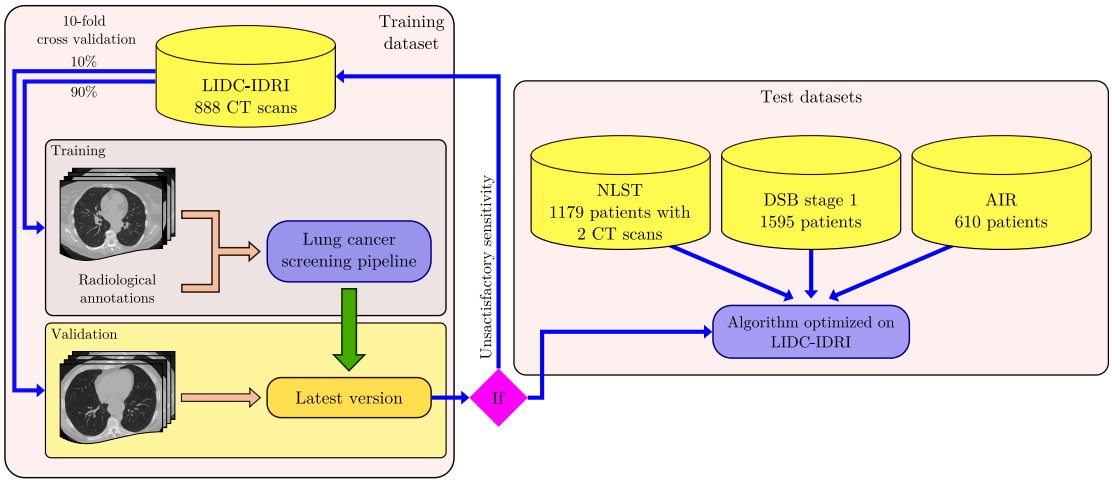


Fig. 6.5: Training and testing approach: the pipeline is trained and validated on the LIDC-IDRI dataset using 10-fold cross-validation, and tested on the NLST dataset, Kaggle and AIR datasets.

3. Nodule detection,

4. Nodule characterization and outcome at the scan level.

A global overview of the pipeline is given in Fig. 6.6.

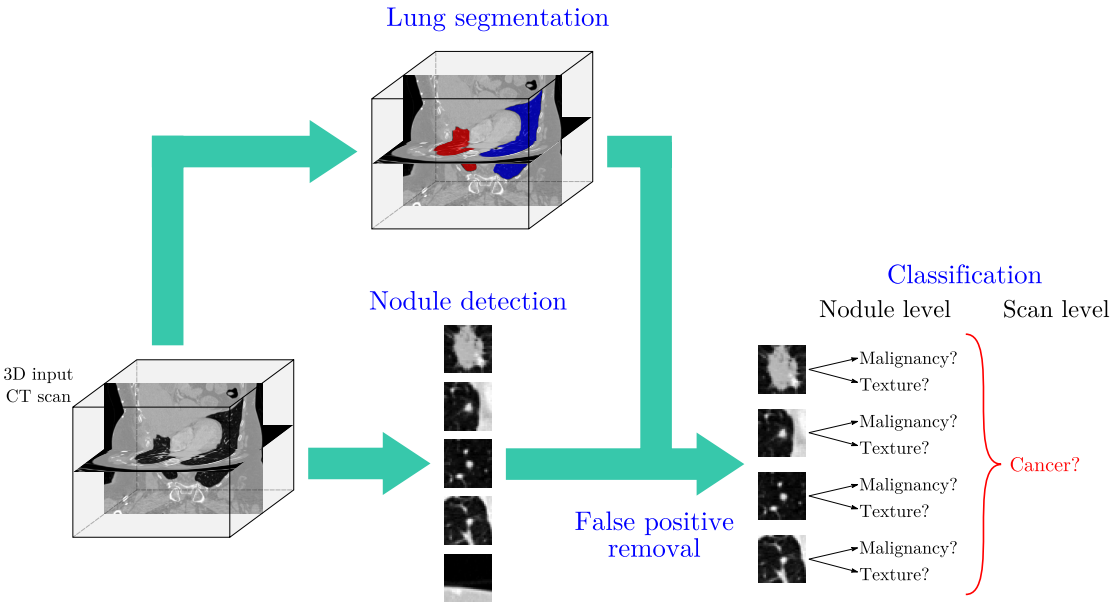


Fig. 6.6: Overview of the lung screening pipeline composed of 3 steps: lung segmentation, nodule detection and nodule characterization.

The image pre-processing step is simple and involves a re-sampling of the raw CT scans to obtain a common spatial resolution of 1 mm in all directions. Intensity values are clipped between -1200 and 600 HU and then linearly transformed to the range $[-1, 1]$.

The second step is the deep learning-based segmentation of the lung parenchyma with a first network whose details are given is section 6.3.2. The segmentation masks of the right and left lungs will be used to remove the false positive nodule candidates located outside the pulmonary region.

Nodules are detected using a second network described in section 6.3.3. Outputs are bounding box coordinates associated with a confidence score. These results are then filtered with the lung segmentation mask.

Finally, each nodule proposal is analysed by a third network presented in section 6.3.4. The characterization includes a texture classification (ground-glass, part-solid or solid) and a malignancy assessment. Results at the nodule level are aggregated in a simple manner to produce an evaluation at the scan level: a scan is labelled as cancer once one nodule is identified as suspicious. The malignancy confidence of the most suspicious nodule is reported as the final cancer probability of the scan.

Our pipeline is thus entirely based on deep learning and includes a total of three networks. The decision rule at the scan level is more straightforward than what was proposed in [Ozdemir et al., 2020; Bonavita et al., 2020], where the k most suspicious nodules are fed into a final classifier predicting cancer at the patient level. The latter strategy imposes to have reliable scan labels, which is not the case for the LIDC-IDRI dataset. Their pipeline is also more complex and the outcome less intuitive and interpretable than our decision rule, solely based on the predictions of the characterization network.

6.3.2 Lung segmentation network

Lung parenchyma segmentation is performed in 2D with a network presenting a classical U-net architecture [Ronneberger et al., 2015], depicted in Fig. 6.7. Inputs are 2D axial slices resized to 240×240. At test time, the network outputs pixel-wise probabilities of belonging to one of the three following classes: background, left lung or right lung.

The network is trained with a cross entropy loss which penalizes the output of the final layer performing the classification:

$$Loss = -\frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \delta(G_n, k) \log p_n^k \,, \tag{6.1}$$

where N is the total number of pixels, K is the number of classes, and δ denotes the Kronecker delta. p_n^k is the classification softmax score for pixel n. The cross entropy loss is computed with respect to the ground truth lung mask G provided by the LUNA16 challenge. These masks were themselves obtained using an automatic segmentation algorithm proposed in [Rikxoort et al., 2009] and may contain errors. Therefore, we stress that this first network is not intended to produce the most accurate segmentation, but rather an acceptable enough mask in order to be able to filter out the false positive nodule candidates located outside the lungs.

Weights are initialized at random and the network is trained for 50 epochs with stochastic gradient descent (SGD). Initial learning rate is set to 0.001, batch size to 50, momentum to 0.9 and weight decay to 0.0001. The learning rate is decreased to 10^{-4} and 10^{-5} after 35 and 45 epochs, respectively.

During the training phase, 50 axial slices of every scan in the dataset are seen by the network at each epoch. Slices are selected randomly, with larger weights given to those

located at the base or the apex of the scan. We also use data augmentation, with elastic deformation and scaling.

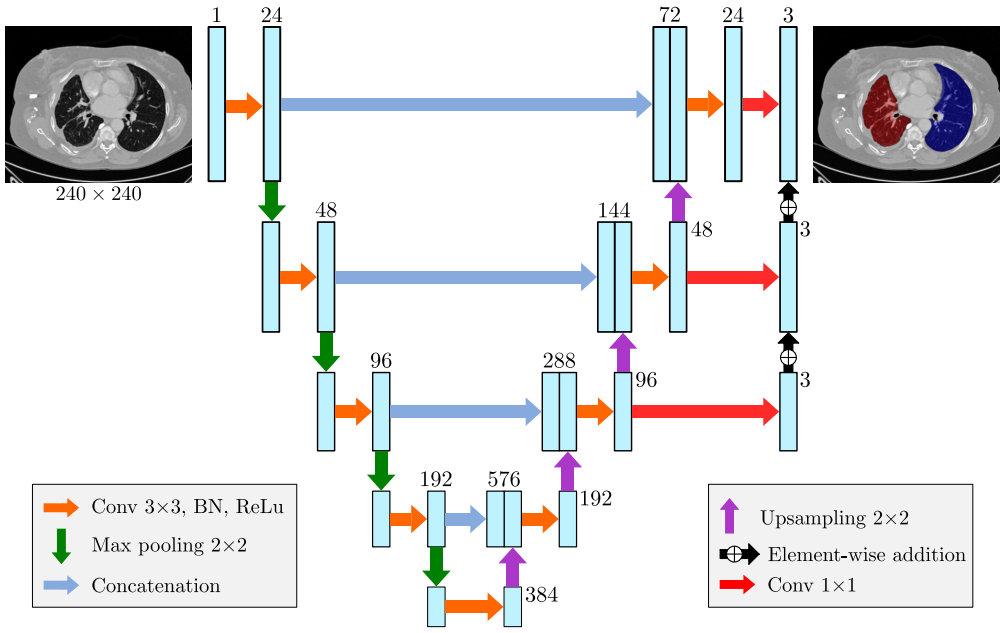


Fig. 6.7: Lung segmentation network. A softmax function is applied after the final layer to produce pixel-wise probabilities corresponding to the three following classes: background, left lung and right lung.

6.3.3 Nodule detection network

Deep learning-based object detectors can be divided in two categories depending on the number of steps of the process [Zhao et al., 2019]. If two stages approaches achieved in the first place better performances, it has also been shown that one stage detectors can lead to competitive results while being faster and simpler to train [Jaeger et al., 2018]. In this chapter, we chose to implement a one stage nodule detection network, inspired by the work of the DSB winning team [Liao et al., 2019]. We followed the same training procedure but we modified the network architecture, shown in Fig. 6.8 and 6.9, which led to improved performances.

The network of the DSB winning team, denoted as DSBWT, had a U-net-like architecture with an encoder-decoder path involving 4 max pooling layers and 2 deconvolution layers. In contrast, the number of max pooling layers is reduced to 2 in our approach, and no up-sampling is performed at the end of the network. Moreover, with a large number of shortcuts, our architecture leverages the concept of skip connections introduced in the residual networks [He et al., 2016], that prevent the problem of vanishing gradients.

The detection is performed in 3D and a patch-based approach is implemented to cope with the GPU memory limitation. Inputs are patches of size $64 \times 96 \times 96$, cropped such that they contain at least one nodule in 70% of the cases. Otherwise, they are centered around a location selected at random in the CT scan. Moreover, the LIDC-IDRI dataset is imbalanced with respect to the nodule diameter: small nodules are over-represented in comparison to larger ones. We mitigate this issue by multiplying by 2 and 6 the sampling frequencies of nodules larger than 30 and 40 mm, respectively.

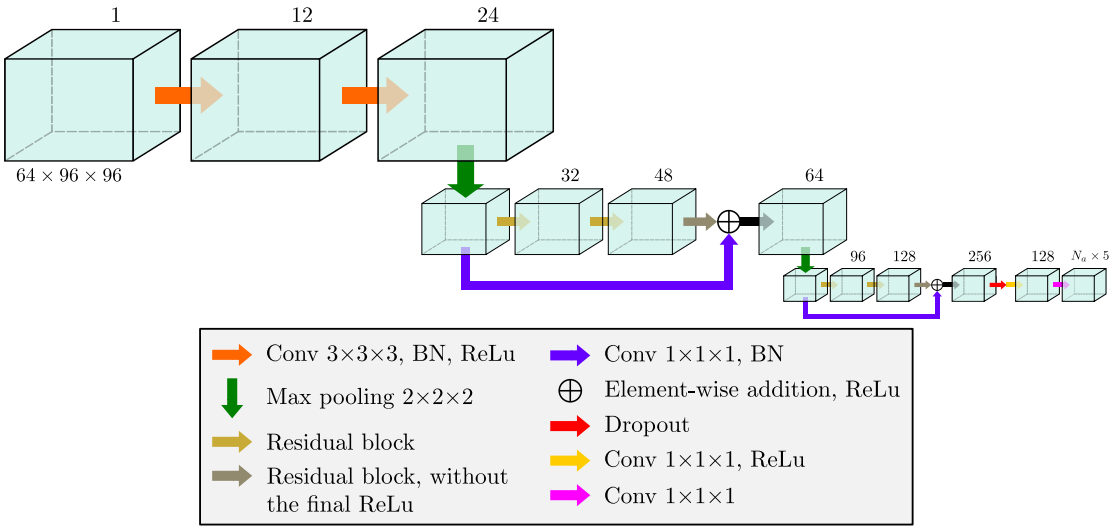


Fig. 6.8: Nodule detection network. Outputs are bounding box coordinates associated with a confidence score. $N_a = 3$ is the number of anchors centered at each voxel in the output cube.

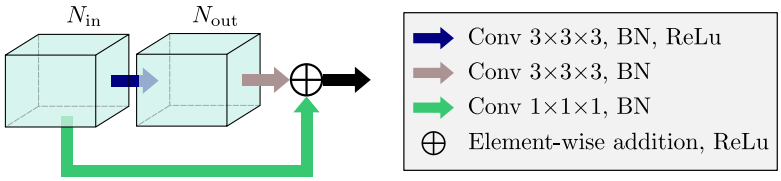


Fig. 6.9: Architecture of a residual block.

The last layer of the network is similar to those of the region proposal networks (RPN) [Ren et al., 2017]. It performs nodule detection based on a dictionary of cubic reference bounding boxes, also denoted as anchor boxes, characterized by their position $A_{x,y,z}$ and their side length A_l. In practice, 3 anchors with side lengths of 10, 30 and 60 mm are centered at each voxel of the output map. The multi-scale approach allows nodules of different sizes to be detected.

For each anchor, the network predicts a confidence score, $\hat{p}$, and 4 location parameters, $\hat{t}_x$, $\hat{t}_y$, $\hat{t}_z$ and $\hat{t}_l$, leading to an output map of size $16 \times 24 \times 24 \times 3 \times 5$. A sigmoid function is applied to the confidence score, such that it represents the probability that the predicted position matches with that of a nodule. The location parameters are an estimate of the offset between the anchor and the ground truth bounding box, G.

During training, the classification label $p \in \{0, 1\}$ of each anchor is defined by computing the intersection over union (IoU) with the ground truth bounding boxes of the nodules. Anchors with an IoU result larger than 0.5 are considered as positive, with $p = 1$. In contrast, negative samples, with $p = 0$, present an IoU score lower than 0.2. Other anchors are not taken into account in the training process.

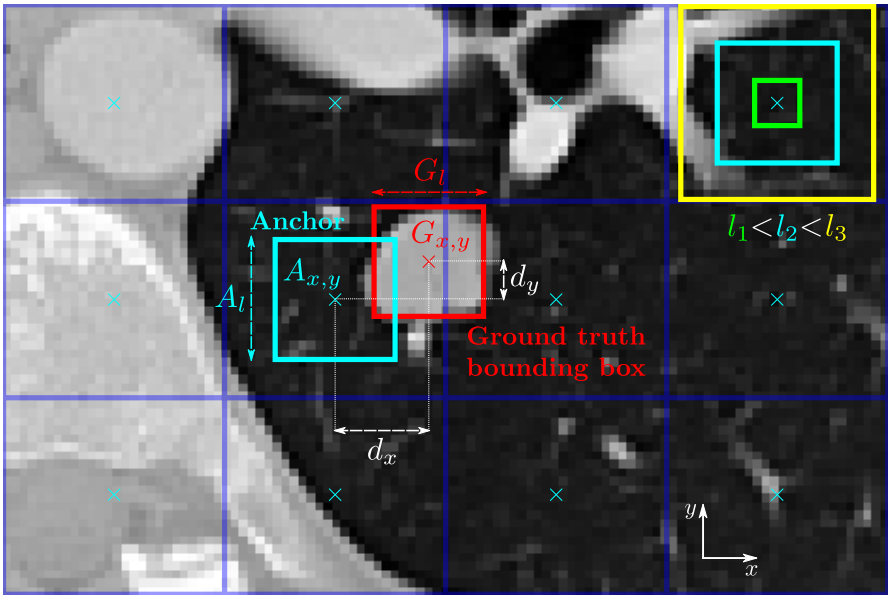


Fig. 6.10: 2D representation of the anchor framework. The output feature map grid of size 3×4 is superimposed over an example input image. 3 anchors of different sizes are centered at each output pixel but are shown only once for clarity. The network objective is to learn a scale-invariant transformation of the offset between anchors and the ground truth bounding boxes. In our network, this approach is implemented in 3D.

We follow [Girshick et al., 2014] for the parametrization of the regression targets t_x, t_y, t_z and t_l, which leads to:

$$t_x = (G_x - A_x)/A_l\,, \tag{6.2}$$
$$t_y = (G_y - A_y)/A_l\,, \tag{6.3}$$
$$t_z = (G_z - A_z)/A_l\,, \tag{6.4}$$
$$t_l = \log(G_l/A_l)\,, \tag{6.5}$$

where G_x, G_y and G_z are the coordinates of the nodule ground truth bounding box, and G_l its side length.

The training loss is composed of two parts: a binary cross entropy loss for the classification term and a smooth L_1 penalty for the regression parameters. It is written:

$$\begin{aligned} Loss = -\frac{1}{NN_a}\sum_{n=1}^{N}\sum_{a=1}^{N_a}\Big[& p_{na}\log\hat{p}_{na} + (1-p_{na})\log(1-\hat{p}_{na}) \\ & - p_{na}\sum_{s\in\{x,y,z,l\}} d\left(t^s_{na}, \hat{t}^s_{na}\right)\Big]\,, \end{aligned} \tag{6.6}$$

where N is the number of voxels of the output map, N_a is the number of anchor side lengths, and $d(x, y)$ is the smooth L_1 loss defined as:

$$d(x,y) = \begin{cases} \frac{(x-y)^2}{2}, & \text{if } |x-y| < 1. \\ |x-y| - \frac{1}{2}, & \text{otherwise.} \end{cases} \tag{6.7}$$

In object detection, the number of negative samples is typically higher than the number of positive ones, which leads to difficulties in training the network. This is particularly true for nodule detection, as they represent a very small fraction of the CT volume. To cope with this imbalance, a hard negative mining approach is implemented, following [Liao et al., 2019]. When computing the loss, the negative anchors are sorted according to their confidence score, and only the top-k candidates are taken into account, the others being discarded. These anchors, which present the largest confidence scores, are denoted as the hard negatives. They correspond to the negative positions which confuse the most the network, because of their resemblance to nodules.

Weights are initialized randomly and the model is trained with a SGD optimizer for 150 epochs with an initial learning rate of 0.01 and a batch size set to 25. The momentum and weight decay are set to the same values as in previous section. The learning rate is reduced linearly to reach the value 0 at the final iteration. Image scaling and rotations are used to perform data augmentation during training.

6.3.4 Characterization network

In [Kim et al., 2019], authors introduced a network architecture for nodule candidate false positive reduction showing competitive results. We re-implemented the same network with some small modifications to perform nodule texture and malignancy classification. An overall representation of the network is presented in Fig. 6.11. A strength of the approach proposed by [Kim et al., 2019] is the multi-scale view of the network. Several patches of different sizes are cropped around the nodule candidate center location and fed into the network in a hierarchical fashion, as shown in Fig. 6.12. Thus, the network is able to capture sharp details while also taking into account larger range contextual information.

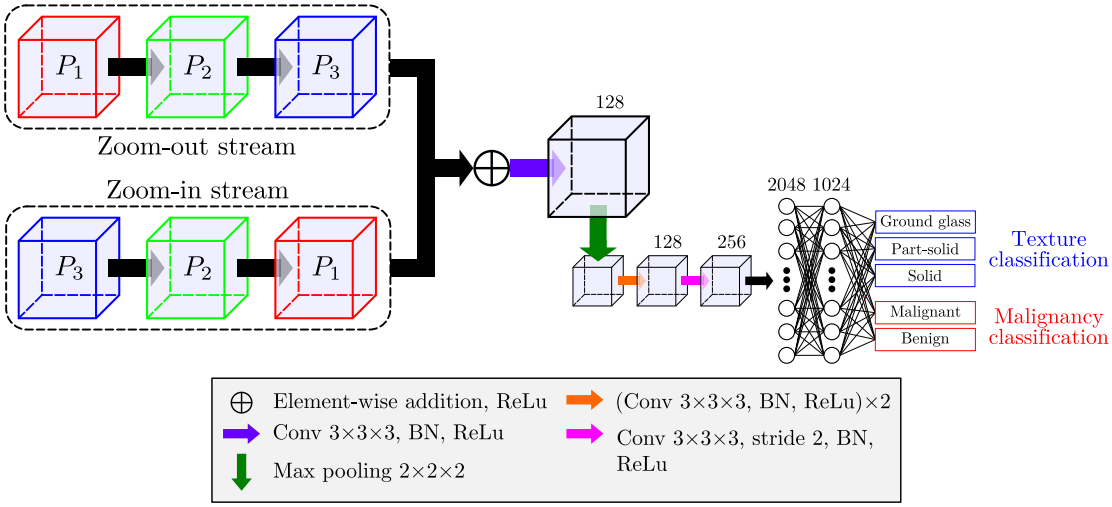


Fig. 6.11: Characterization network. A softmax and sigmoid functions are applied after the final fully connected layer to generate the texture and malignancy probabilities, respectively.

In comparison with the original approach, we modified the input patch sizes and parameters of the layers to perform the convolutions in 3D. Moreover, instead of a unique

binary output, the network now predicts the probabilities of belonging to one of the tree texture classes, together with a malignancy assessment.

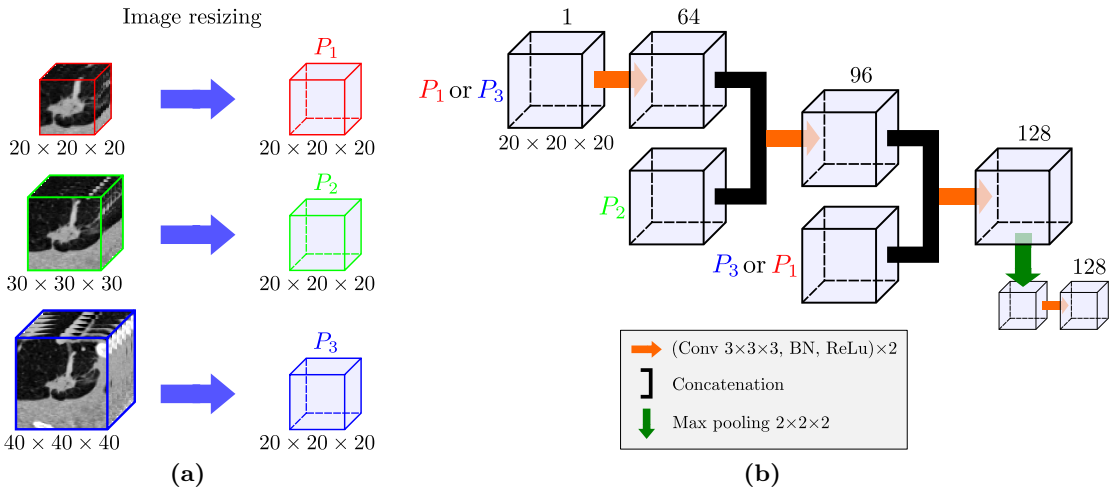


Fig. 6.12: Inputs of the characterization network (6.12a) and architecture of the zoom-in and zoom-out streams (6.12b).

The training dataset is composed of the LIDC-IDRI nodule ground truth bounding boxes plus false positive candidates generated by the detection network on its training data. The network is trained with 2 cross entropy losses penalizing the malignancy and texture scores produced by the final layer of the network:

$$Loss = -\frac{1}{N}\sum_{n=1}^{N}\Big[\alpha\, m_n \log \hat{m}_n + (1-m_n)\log(1-\hat{m}_n) + \lambda_n \sum_{k=1}^{3} w_k \delta(T_n, k)\log p_n^k\Big], \tag{6.8}$$

where N is the total number of nodule locations of the training set and δ denotes the Kronecker delta. m_n and $\hat{m}_n$ are the ground truth malignancy label and the network prediction for the n-th bounding box, respectively. The detection false positives used to extend the training data are labelled as non cancer, i.e., $m = 0$. Furthermore, they are not taken into account in the texture loss computation by setting $\lambda = 0$. This coefficient is otherwise set to 1 for the LIDC-IDRI ground truth bounding boxes. The texture loss is computed between the ground truth texture label, denoted as T, and the classification softmax score p^k produced by the network.

Ground-glass, part-solid and malignant nodules are under-represented in the training data. A solution to alleviate this issue is to compute a weighted cross entropy loss, by the introduction of the parameters α and w for the malignancy and texture classifications, respectively. These parameters are set such that all classes contribute equally to the loss function.

An SGD optimizer is again used to train the network for 80 epochs. The initial learning rate is set to 0.003 and reduced linearly to reach the value 0 at the final iteration. Weights are initialized as previously and the momentum and weight decay remain unchanged.

6.4 Results

6.4.1 Cross-validation results on the LIDC-IDRI dataset

In this section, we perform an end-to-end analysis of our lung screening pipeline on the LIDC-IDRI dataset based on a 10-fold cross-validation procedure. Results presented here are derived from the aggregation of those obtained on each of the 10 folds.

Lung segmentation.

The first step of the pipeline is the segmentation of the lung parenchyma, whose mask is used afterwards to remove false positive candidates. Two examples of results generated by the 2D U-net are presented in Fig. 6.13, together with the segmentation provided by the LUNA16 challenge. The latter was used as a surrogate ground truth to train the network, although it was itself generated by an automatic algorithm and contains errors, for instance visible in the second example in Fig. 6.13.

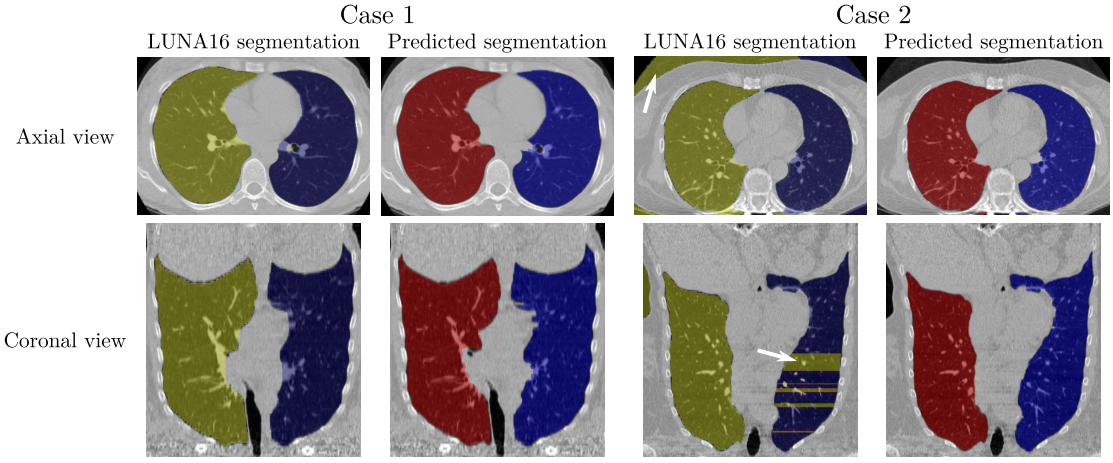


Fig. 6.13: Visualization of some lung segmentation results. The second example is a case where the LUNA16 segmentation contains errors, highlighted by the arrows.

Fig. 6.14 shows quantitative results computed with respect to the LUNA16 segmentation. They are presented for information purpose only as the LUNA16 segmentation cannot be considered as ground truth. Nevertheless, we obtain median Dice scores above 0.98 for both lungs.

Nodule detection.

Fig. 6.15 summarizes the detection results obtained with 10-fold cross-validation on the LIDC-IDRI database, after removal of false positives located outside the lung region using the segmentation mask generated in the previous section. A candidate bounding box is considered as a true positive if its center of coordinates (C_x, C_y, C_z) falls within the sphere centered at the nodule ground truth location, i.e., if $\sum_{s\in\{x,y,z\}}(C_s - G_s)^2 < G_l^2$, where (G_x, G_y, G_z) and G_l are the nodule ground truth bounding box coordinates and its side length, respectively. This decision rule is the same as the one used in the LUNA16 challenge.

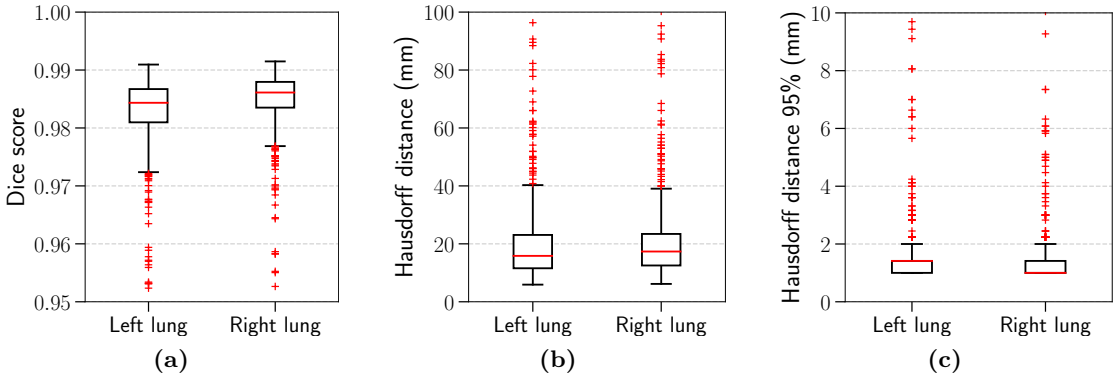


Fig. 6.14: Quantitative segmentation results with respect to the LUNA16 segmentation.

1138 nodules were correctly identified out of 1186 annotated by at least 3 radiologists. 8912 locations were wrongly extracted by the network, leading to a sensitivity of 96% for 10 false positive candidates per scan in average. Examples of true positive bounding boxes are shown in Fig. 6.16, and Fig. 6.17 presents false positive candidates randomly selected.

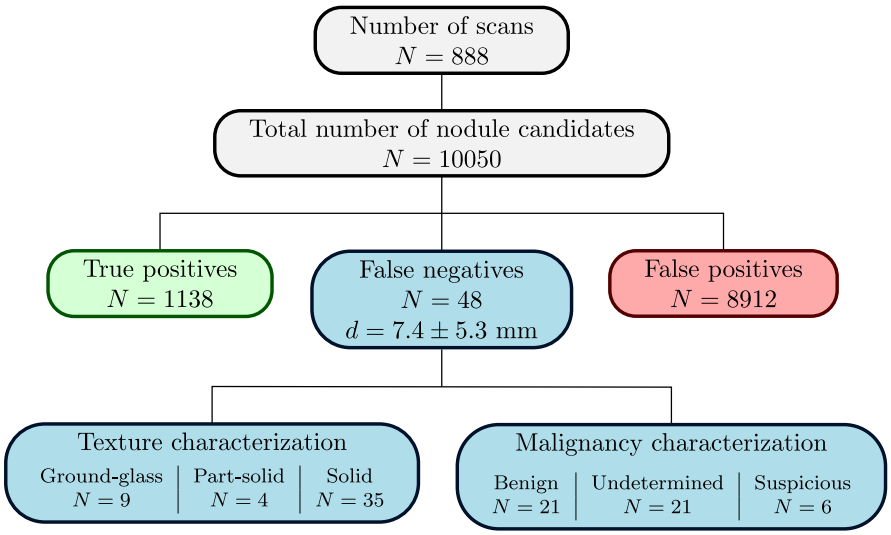


Fig. 6.15: Nodule detection results obtained on the LIDC-IDRI dataset with 10-fold cross-validation.

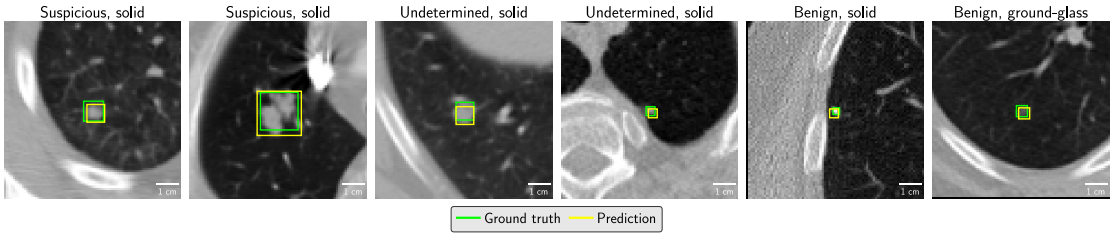


Fig. 6.16: Example of 6 nodules correctly identified by the nodule detector with their texture and malignancy characteristics.

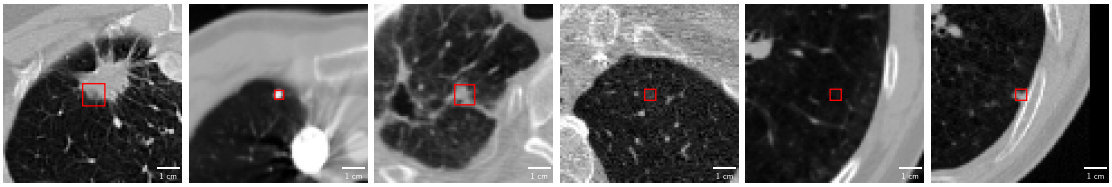


Fig. 6.17: Visualization of 6 false positive candidate locations randomly selected.

The network missed 48 nodules, 6 of whom were identified as suspicious by the radiologists and are shown in Fig 6.18, top row. The mean diameter of the false negatives is 7.4 mm.

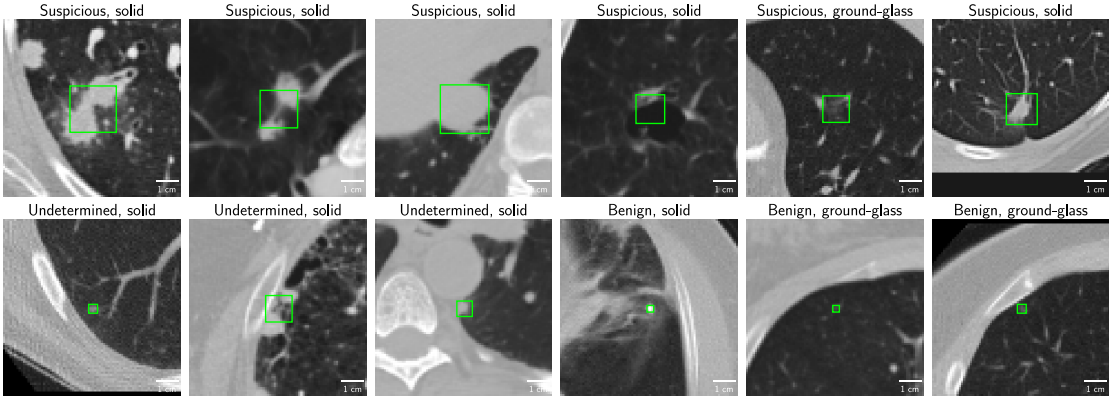


Fig. 6.18: Visualization of the 6 suspicious false negatives (top row), together with 3 benign and 3 undetermined nodules randomly selected among the missed lesions (bottom row).

The free response operating characteristic (FROC) curve is the counterpart of the ROC curve for object detection. Variations of the sensitivity of the detector with respect to the average number of false positive per scan are obtained by changing the detection threshold. The FROC curves of the DSBWT baseline and of our network are plotted in Fig. 6.19a. This figure takes into account all nodule candidates extracted by the detector. A different approach was followed in the LUNA16 challenge, which established a list of excluded annotations regrouping non-nodules, nodules of size below 3 mm or annotated only by one or two radiologists. Candidates matching excluded annotations are then removed from the analysis. The consequence is a decrease in the number of false positives, as shown in Fig. 6.19b. From an object detection viewpoint, these findings are not totally irrelevant as they were identified by at least one radiologist, which explains the approach proposed by the LUNA16 challenge [Setio et al., 2017]. However, they have to be taken into account when analysing a whole pipeline, which is the aim of this chapter.

In comparison to the DSBWT network, our approach achieves better performances with higher sensitivities, thus demonstrating the relevance of our proposed architecture.

Nodule characterization.

In this section, nodule candidates identified by the detector are characterized by the last network in terms of texture and malignancy. True nodule locations missed by the

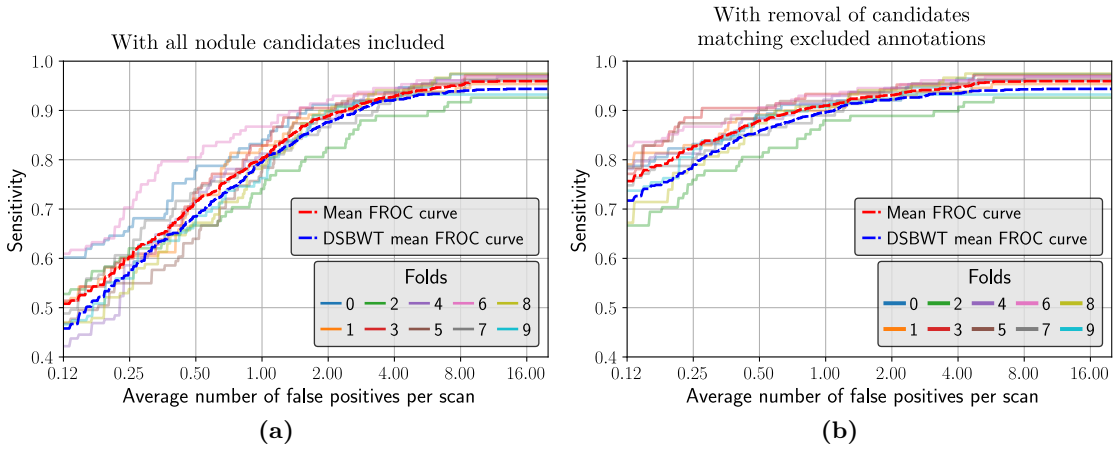


Fig. 6.19: FROC curve results, considering all nodule candidates (6.19a) or removing those matching excluded annotations, as proposed by the LUNA16 challenge (6.19b).

detection network are not taken into account in the results presented in this section, as they are not seen by the characterization network. However, they will be considered for the result aggregation at the scan level presented in the next section.

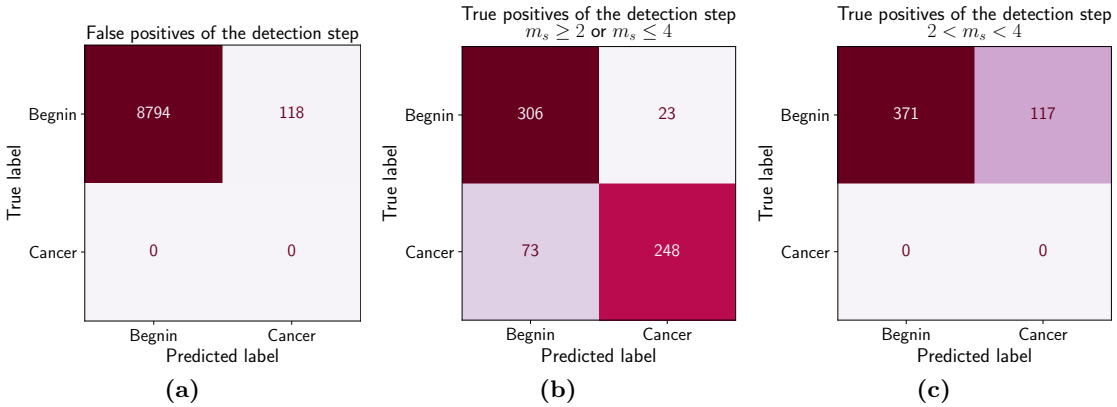


Fig. 6.20: Confusion matrices (CM) for the malignancy characterization of the nodule candidates. (6.20a) CM of the false positives of the detection step. (6.20b) CM of the true positives of the detection step, which were identified as benign of suspicious by the radiologists. (6.20c) CM of the true positives of the detection step with an ambiguous malignancy score. Their true label is set to benign arbitrarily.

Fig. 6.20 presents the confusion matrices obtained for the malignancy assessment. False positives of the detection step are wrong nodule locations proposed by the detector and 99% of them are correctly classified as benign by the characterization network. We then separated the results between true positives of the detection step with a reliable malignancy score and those with an ambiguous one. The ground truth label for the latter is arbitrarily set to benign. Because of their unreliable label, these nodules were not taken into account when computing the ROC curve and the metrics shown in Fig. 6.21a and Tab. 6.2, respectively. The good AUC score has to be qualified because of the class

imbalance and the large number of true negatives. False negative and positive examples are presented in Fig. 6.22 and 6.23, respectively.

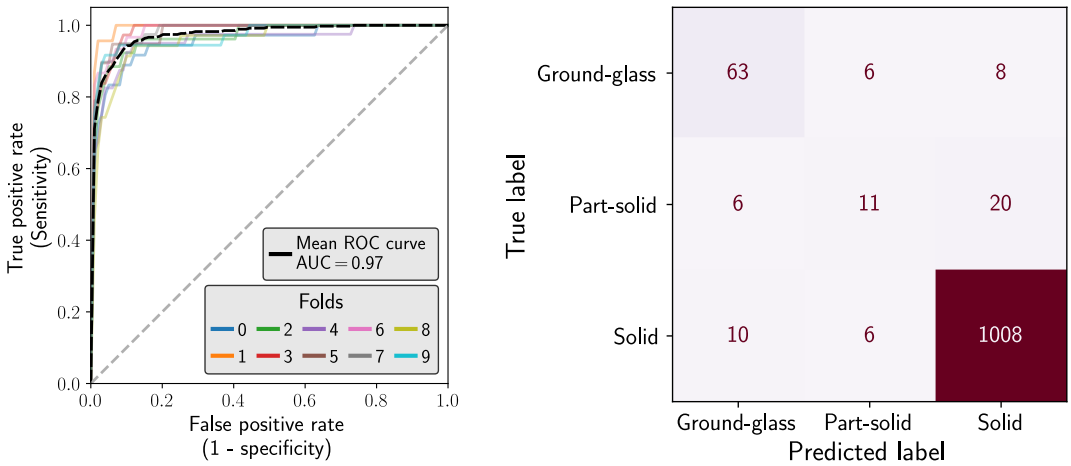


Fig. 6.21: (6.21a) ROC curve of the nodule candidate characterization. Nodules with ambiguous malignancy ground truth label are excluded from the analysis. (6.21b) Confusion matrix for the texture classification of the true positives of the detection step.

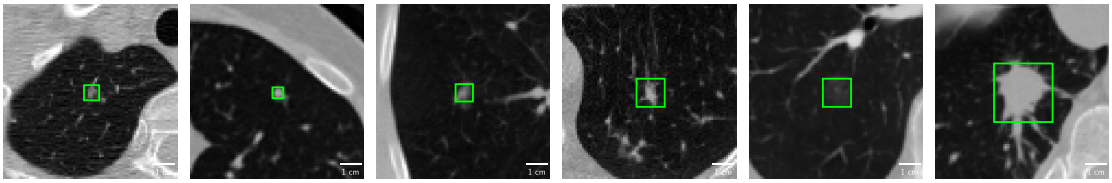


Fig. 6.22: Example of 6 suspicious nodules wrongly classified as benign.

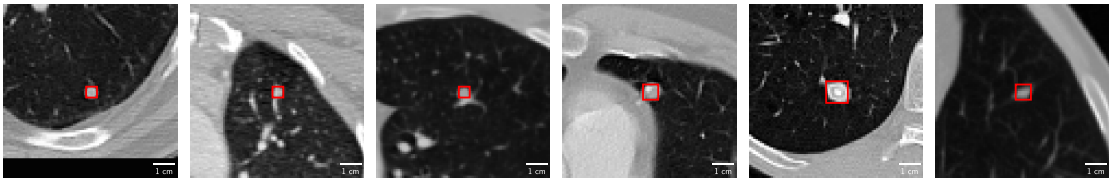


Fig. 6.23: Example of 6 benign locations wrongly classified as suspicious.

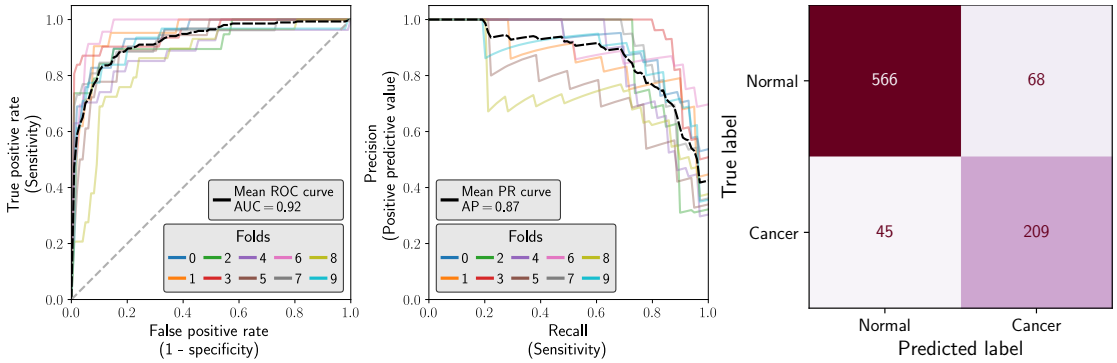
In addition to providing a malignancy assessment, the second objective of the characterization network is to describe the lesion appearance. The confusion matrix related to this characterization is presented in Fig. 6.21b. False positives of the detection step are excluded from the analysis. Most of the nodules were annotated as solid by the radiologists and are correctly identified as such by the network. However, despite the class imbalance, 80% of the ground-glass nodules are rightly classified. Metrics corresponding to the texture assessment are presented in Tab. 6.2.

Tab. 6.2: Results of the characterization of the candidate lesions identified by the nodule detector, obtained by 10-fold cross-validation on the LIDC-IDRI dataset. Nodules with ambiguous malignancy score are removed from the malignancy analysis. Texture results are obtained by computing the metrics globally over the 3 classes.

	Precision (%)	Recall (%)	F1-score (%)
Malignancy	63.8 ± 4.9	78.0 ± 8.7	69.9 ± 5.0
Texture	95.1 ± 1.6	95.1 ± 1.6	95.1 ± 1.6

Prediction at the patient level.

Finally, the last step of the pipeline is a cancer assessment at the scan level. This is done simply in our case by aggregating results at the nodule level according to the rule stated in section 6.3.1.



(a) **(b)** **(c)**

Fig. 6.24: Results at the patient level, in terms of ROC curve (6.24a), precision-recall curve (6.24b) and confusion matrix (6.24c).

The confusion matrix, precision-recall (PR) and ROC curves associated to the prediction at the scan level are visible in Fig. 6.24. 209 cases out of 254 are correctly identified as cancer, with 68 false positives. Performance metrics are given in Tab. 6.3: the precision for the subject cancer classification is of 75%, and the recall of 82%.

Tab. 6.3: Cancer classification results at the patient level obtained with 10-fold cross-validation on the LIDC-IDRI dataset.

Precision (%)	Recall (%)	F1-score (%)
75.4 ± 7.8	82.6 ± 7.4	78.4 ± 5.1

6.4.2 Tests on independent datasets

The main limitation of the LIDC-IDRI database is the absence of nodule malignancy labels confirmed by biopsy, raising the question of the performances of the pipeline on images for which the patient cancer status is available. We investigate this issue by applying our pipeline trained on the full LIDC-IDRI database on the three independent test sets presented in section 6.2. In addition, these datasets allow the generalisation

capabilities of the pipeline to be tested, particularly to images corresponding to different populations and obtained by different devices.

Results on the NLST dataset

In a first experiment, we apply the detection network on the T_2 image of the cancer-free subjects, and on the image at diagnosis for the cancer patients. On this set of 1179 images, 2352 nodules were annotated by the radiologists, including 177 malignant lesions. The annotations of the radiologists include localization informations, but no details regarding the size of the lesions are provided. Thus, we modify the detection decision rule: a candidate bounding box is now considered as positive if $\sum_{s\in\{x,y,z\}}(C_s - G_s)^2 < d^2$, where d is a distance threshold.

With a 1 cm threshold, our system detected 75% of all annotated nodules, including 1730 benign nodules out of 2352 (73%), and 170 malignant nodules out of 177 (96%), for 12 false positives per scan in average. With a 3 cm threshold, the sensitivity is increased with 172 malignant nodules detected (97%). The 5 undetected lesions were located next to the mediastinum.

In a second step, we apply our detection algorithm on the CT scan collected one year before the diagnosis for the cancer patients. Out of the 177 malignant lesions, 20 are not visible one year before. Among the 157 already visible lesions, 152 (97%) were successfully detected by our network.

The characterization of the detected nodules and the predictions at the scan level were still ongoing at the time of writing. Therefore, no results related to these steps of the pipeline on NLST can be presented yet.

Results on the DSB stage 1 dataset

In contrast to the NLST dataset, annotations regarding the localization and the characterization of nodules are not provided in the DSB stage 1 dataset. The only available information is the cancer status at the scan level, which prevents any evaluation of the detection and characterization steps of the pipeline.

The predictions of the pipeline at the subject level for the DSB stage 1 data are presented in Fig. 6.25. This dataset contains 1595 CT scans, and 419 of which are cancer patients. 298 (71%) cancer cases are successfully identified by our system. 283 cancer-free subjects out of 1176 (24%) are wrongly classified as positive. The AUC and average precision scores are 0.8 and 0.58, respectively. Performance metrics are given in Tab. 6.4.

Tab. 6.4: Cancer classification results at the patient level obtained on the DSB stage 1 dataset.

Precision (%)	Recall (%)	F1-score (%)
51.3	71.1	59.6

Results on the AIR cohort

As for previous section, the cancer status of the subject is the only available information in the private database of COPD patients. Predictions results at the scan level are

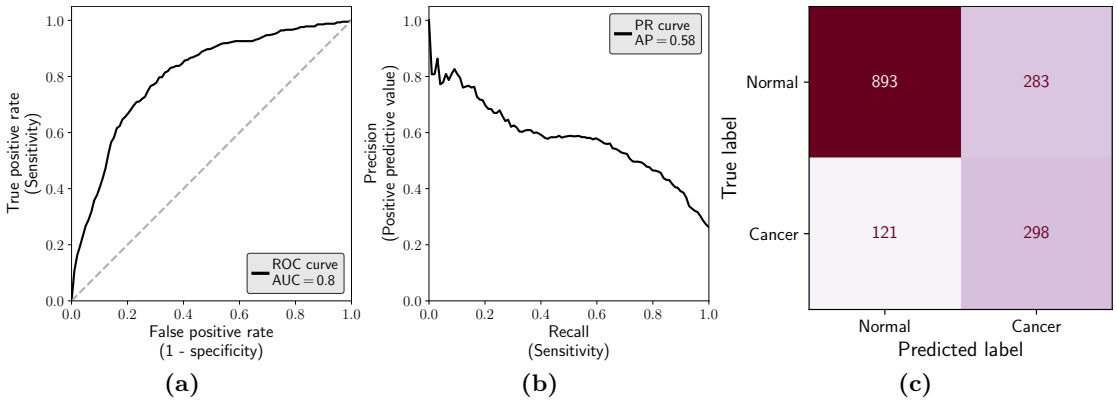


Fig. 6.25: ROC curve (6.25a), precision-recall curve (6.25b) and confusion matrix (6.25c) for the predictions at the patient level on the DSB stage 1 dataset.

presented in Fig. 6.26. Among the 610 subjects of the cohort, 22 were diagnosed with cancer. 13 (59%) are correctly detected, and 125 cancer-free subjects out of 588 (21%) are miss-classified by the pipeline. The AUC and average precision scores are 0.78 and 0.14, respectively. Performance metrics are given in Tab. 6.5.

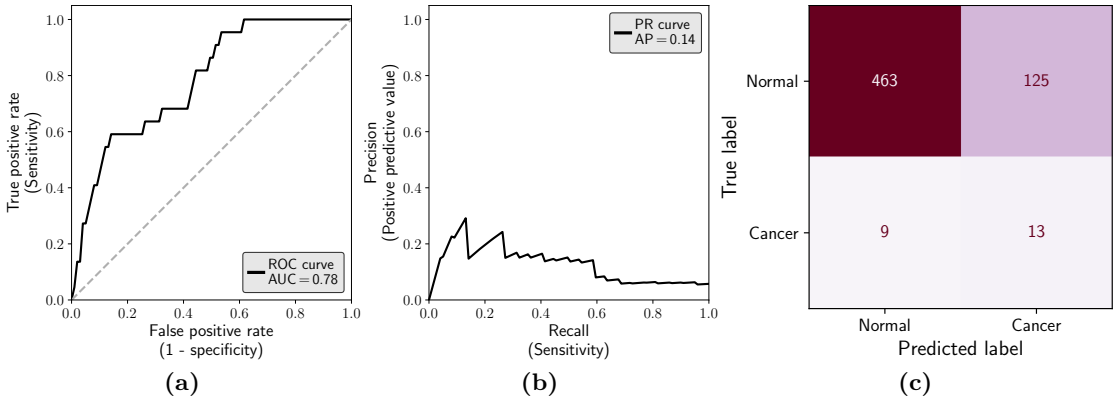


Fig. 6.26: ROC curve (6.26a), precision-recall curve (6.26b) and confusion matrix (6.26c) for the predictions at the patient level on the AIR dataset.

Tab. 6.5: Cancer classification results at the patient level obtained on the AIR dataset.

Precision (%)	Recall (%)	F1-score (%)
9.4	59.1	16.2

Tab. 6.6 summarizes the performances of the pipeline obtained on the LIDC-IDRI dataset by cross-validation and on the test sets.

6.5 Discussion

One of the main contributions of this chapter is the evaluation up to the patient level of a lung cancer screening pipeline. Direct comparison with previous works is difficult

Tab. 6.6: Performance results (in %) for the cancer prediction at the patient level on the LIDC-IDRI dataset and on the test sets. In all cases, the pipeline was solely trained using LIDC-IDRI data. Results for the NLST subset were not yet available at the time of writing. (CV: cross-validation.)

	LIDC-IDRI (10-fold CV)	DSB stage 1	AIR
Precision	75.4 ± 7.8	51.3	9.4
Recall (Sensitivity)	82.6 ± 7.4	71.1	59.1
F1-score	78.4 ± 5.1	59.6	16.2
Specificity	89.3 ± 4.0	75.9	78.7

because of the variations in data sources, in training and validation procedures, and in the quality of annotations. Regarding the predictions at the patient level, [Zhang et al., 2019] reported a recall score of 84% obtained by 10-fold cross-validation on a proprietary dataset, but the precision score is not given. [Bonavita et al., 2020] obtained on a small subset of the LIDC-IDRI dataset a precision and recall scores of 84% and 80%, respectively, but these results were not validated on external datasets. [Liao et al., 2019] and [Ozdemir et al., 2020] achieved both a 0.87 AUC on the DSB stage 2 dataset, but their networks were not trained only on the LIDC-IDRI dataset, but also on the DSB stage 1 data. A 0.94 AUC was obtained by [Ardila et al., 2019], with a system trained and tested on the NLST database. Therefore, our results are in the same order of magnitude as those reported in previous works.

One can note that all systems, including ours, achieve good, even excellent, AUC scores. These results might however be misleading due to the heavily imbalanced nature of the data. Cancer-free subjects are indeed much more numerous in all datasets than lung cancer patients, which represent only 30 and 15% of the subjects in the DSB stage 1 and NLST datasets, respectively. In case of such imbalance, it has been reported that scores like accuracy and AUC tend to be over optimistic with respect to the prediction performances [Davis & Goadrich, 2006]. This can be explained by the fact that a naive approach classifying correctly all negative subjects would already lead to good accuracy results. In contrast, recall (a.k.a. the sensitivity) and precision focus on the predictions related to the minority class, which is often the group of interest, in particular in the medical domain [He & Garcia, 2009].

The discrepancies between the AUC scores and the precision-recall results are clearly visible when comparing the ROC and PR curves obtained on the DSB stage 1 data, shown in Fig. 6.25. The ROC analysis leads to a rather good AUC score of 0.8, while the average precision result is poorer (0.58), but reflects certainly better the real performances of the pipeline. Comparison of these results with previous work is difficult as theses metrics are often not provided. In particular, the excellent AUC scores reported by [Ardila et al., 2019] on the NLST dataset would have benefited from a confirmation by a precision-recall analysis.

The main limitation of the cross-validation results on the LIDC-IDRI dataset is the absence of ground truth cancer status, both at the nodule and scan levels. Moreover, little information is provided for lung masses despite their relevance for lung cancer screening. Furthermore, the LIDC-IDRI annotations are not claimed to be extensive,

and some lesions may have been missed. Yet, the annotation quality has a direct impact on the algorithm training and on its performance evaluation. In addition to the absence of cancer status labels, the LIDC-IDRI annotations exhibit a high level of inter-rater variability. There has been a consensus between 3 or 4 radiologists for only 52% of all annotated lesions. Nodules identified by only 1 or 2 radiologists, so-called irrelevant findings in the LUNA16 challenge, were not taken into account for the training of the pipeline and were then considered as false positives during the evaluation. Yet, they constitute pulmonary abnormalities that could be of significance for lung cancer detection. Interestingly, among the nodules annotated by only 1 or 2 experts, 104 present a malignancy score above 4, and would have thus been considered as suspicious within our framework.

Moreover, the LIDC-IDRI malignancy score is purely based on the subjective evaluation of the image by the radiologists. Yet, in the NLST study, 96.4% of the scans identified as positive turned out to be false positives [NLST, 2011]. Direct extrapolation of this result to the LIDC-IDRI database is not possible, as the NLST scans were analyzed in real life clinical conditions, whereas a two steps annotation procedure was established for the LIDC-IDRI dataset [McNitt-Gray et al., 2007]. Nevertheless, the high false positive rate in the NLST study shows that the malignancy scores need to be considered with caution.

Another important contribution of our work is the evaluation of our pipeline, only trained on the LIDC-IDRI dataset, on three independent test sets, where the subject cancer status is known. At the time of writing, the complete analysis was not not yet available for the NLST dataset. Nevertheless, a part of the DSB stage 1 data is derived from the NLST study. Therefore, results obtained on the DSB stage 1 data provides an insight of the performances that could be achieved on the NLST dataset.

Regarding the detection step of the pipeline, the average number of candidates per scan identified by the detector varies little across the 4 datasets: 12, 14, 13 and 13 on the LIDC-IDRI, the NLST, the DSB stage 1 and the AIR datasets, respectively. The undetected nodules on the NLST were all located next to the mediastinum. This region is particularly challenging with the presence of the bronchii, which could explain the network failure for these cases. Moreover, the second experiment conducted on the NLST data demonstrated the ability of the network to detect malignant nodules one year before diagnosis.

In addition to the comparison with existing computerized screening frameworks, another important question is the performance of the pipeline compared to that of the radiologists. The confusion matrix summarizing the performances of the radiologists in the NLST trial is shown in Fig. 6.27. It is obtained by pooling the results of the LDCT arm of the trial over the three screening rounds, T_0, T_1 and T_2. In the NLST study, a scan was identified as positive if a lesion above 4 mm was detected, or any other lung abnormalities relevant for lung cancer. 75126 screenings were performed over the three rounds of the trial. 18146 were marked as positive, but only 649 were confirmed as lung cancer cases. In addition, 44 false negatives were missed by the radiologists. The corresponding precision, recall, and F1 scores are 3.6%, 93.7% and 6.9%, respectively. Thus, the screening by the radiologists in the NLST study exhibits a good recall, but also a high number of false positives. Yet, beside detecting all cancer cases, keeping the false positive rate low is essential in screening. False positives lead to additional costs, unnecessary anxiety for the patient, and potential increased morbidity due to the screening side effects [Patz et al., 2014; USPSTF, 2021]. Specificity results of the

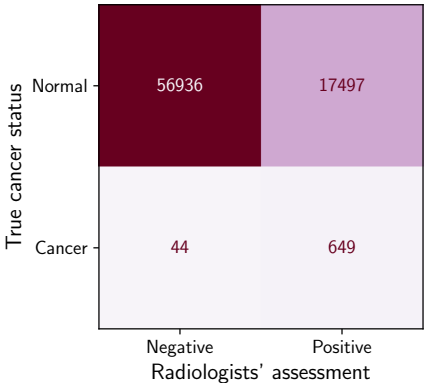


Fig. 6.27: Confusion matrix describing the performances of the radiologists in the NLST study. Results were pooled over the three screening rounds T_0, T_1 and T_2. These results do not take into account cancer cases either identified after the end of the screening period or diagnosed in participants who missed the screening.

radiologists on the NLST study and of our pipeline on the DSB stage 1 dataset are 76.5% and 75.9%, respectively, while the sensitivities are 93.7% and 71.1%. Therefore, our pipeline achieves a specificity close to that of the radiologists, but a lower sensitivity, even though the differences in image sources and fraction of cancer cases between the two datasets limit the comparison. Nevertheless, it shows that the performances of our pipeline are encouraging, in particular given all the limitations related to the training database, mentioned above. One might expect improved performances after re-training on data with higher quality labels, but this will need to be confirmed by future work.

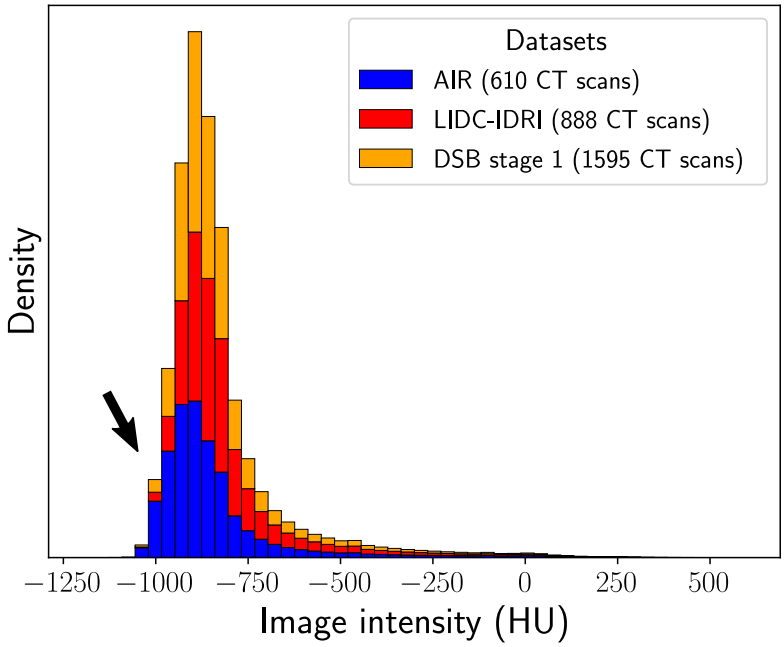


Fig. 6.28: Normalized histograms of the intensity in the lungs for the AIR, LIDC-IDRI and DSB stage 1 datasets. The AIR images present a larger fraction of low attenuation values, as indicated by the black arrow.

Despite close AUC scores, results are poorer on the AIR cohort, in particular for the precision. The percentage of wrongly classified cancer-free subjects (21%) is however similar to the one obtained on the DSB stage 1 data (24%). On the AIR data, the difficulty is thus the identification of the true cancer cases. Several factors might explain the discrepancy with the other datasets. First, the AIR images were collected between 2015 and 2018 whereas all the NLST scans were acquired before 2007 and those of the LIDC-IDRI probably around the same period, as the complete dataset was released in 2011. The AIR database is thus composed of more recent images, certainly acquired with more recent devices, which could be a reason of the drop in performances. It could also be explained by differences in the kernels used for the image reconstruction. Indeed, in computed tomography, images are obtained by processing the raw attenuation data by algorithms [Seeram, 2016]. This reconstruction process involves convolution kernels that have an impact on the amount of noise in the image and therefore on the appearance of the image structures [Neubauer et al., 2016]. In particular, it has been demonstrated that these kernels may alter the performances of downstream algorithms applied on the reconstructed image [Jacobs et al., 2016]. Finally, the AIR cohort corresponds to a specific population with subjects diagnosed with COPD. This disease leads to a progressive deterioration of the lungs and modifies their appearance by increasing the proportion of regions with low attenuation values [Thurlbeck & Müller, 1994]. There is no information regarding the presence of COPD patients in the other datasets. Nevertheless, Fig. 6.28 shows that the average distribution of the intensity in the lungs in the AIR dataset is different to those of the LIDC-IDRI and DSB stage 1 datasets. In particular, the AIR images have clearly a larger number of low attenuation values, as indicated by the black arrow. Thus, the presence of COPD is another possible explanation of the lower performances of the pipeline on the AIR data.

6.6 Conclusion

In this chapter, we provide a complete analysis of an end-to-end lung cancer screening pipeline, fully automated. The framework was trained only on the LIDC-IDRI dataset, for which only limited annotations are available. Nevertheless, it showed compelling results, close to state of the art, on independent test sets. In particular, better results reported in previous works were obtained with pipelines trained on larger datasets with higher quality labels.

Moreover, our study highlights the limitations of lung cancer pipeline comparisons solely based on the AUC score, and demonstrated the relevance of precision-recall analyses.

Our framework, entirely based on deep learning, remains a black box with a lack of interpretability. Yet, interpretability has become a desirable property, especially in the medical domain, where the algorithm output may impact the decisions made about the patient. Making our pipeline more interpretable would be an interesting but difficult challenge, because of the number of networks and steps depending on each other.

The step dedicated to nodule characterization is critical. The NLST study has demonstrated the difficulty of the task, revealed to be challenging even for radiologists. The lack of consensus in the definition of a nodule and the subjectivity of the image-based malignancy assessment are limitations that need to be addressed to improve the performances of automated lung screening pipelines. In particular, annotations

provided only at the scan level limit the training possibilities of the algorithms and prevent the complete understanding of their predictions at test time, when evaluating their performances.

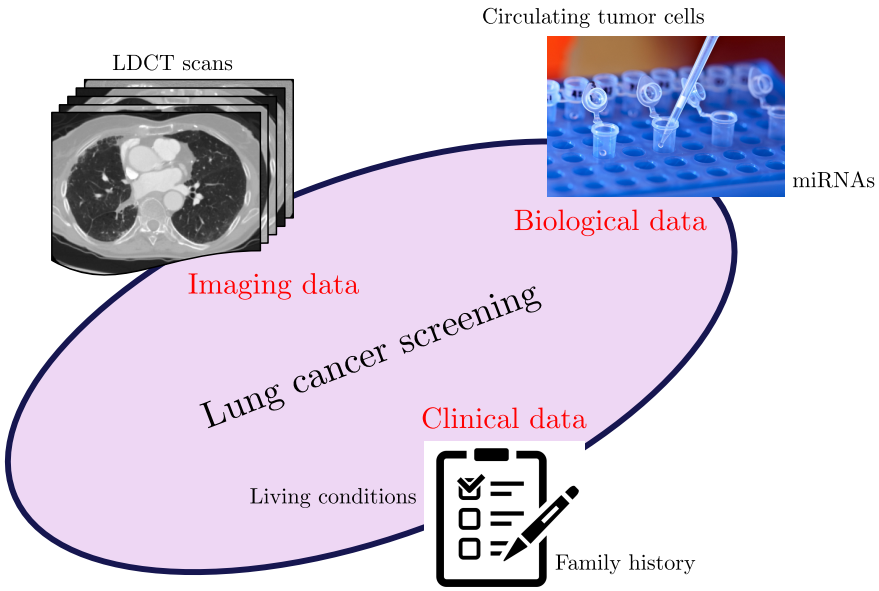


Fig. 6.29: The future of lung cancer screening: combining different data sources.

Besides re-training our pipeline on data with higher quality annotations, better performances may be achieved by taking into account the evolution of the lesions over time. For example, our pipeline could leverage the scans acquired one year apart for each subject in the NLST dataset. Moreover, some works have demonstrated the application of recent advances in image inpainting for synthetic lung nodule generation [Yang et al., 2019; Jin et al., 2018; Kommrusch & Pouchet, 2019; Xu et al., 2019]. These approaches could be explored to augment the training datasets, and to increase in particular the fraction of malignant nodules, in order to alleviate the lack in high quality annotated data.

Finally, one can note that imaging is only one of the numerous approaches explored for lung cancer screening. In particular, there are concerns related to long term radiation exposure in LDCT screening. In contrast, circulant miRNAs, for example, can be easily collected by non-invasive liquid biopsy and their use as a potential biomarker for lung cancer screening is investigated in the appendix of this chapter. In fact, the future of lung cancer screening lies certainly in the combination of different data sources, such as imaging, biological data, and clinical data, as shown in Fig. 6.29. Combining such variable sources would be a challenge regarding the collection, the annotation and the analysis of the data, but is a promising research avenue for future work.

Acknowledgments

This work was partially funded by the French government, through the UCA^{JEDI} and 3IA Côte d'Azur "Investments in the Future" projects managed by the National Research Agency (ANR) with the reference numbers ANR-15-IDEX-01 and ANR-19-P3IA-0002 and supported by the Inria Sophia Antipolis - Méditerranée "NEF" computation cluster.

Chapter 7

Conclusion

Contents

7.1 Main contributions 135
7.2 Perspectives 136
7.2.1 Spatial regularization in neural networks 136
7.2.2 Unsupervised quality control of bounding boxes 137
7.2.3 Consensus estimation in a supervised learning setting 139
7.2.4 Towards an integrative analysis of chest CT scans 141

7.1 Main contributions

We summarize in this section the main contributions of the thesis.

Three of the four main chapters were related to the general topic of probabilistic image segmentation modelling. In chapter 3, we developed some theoretical aspects of the Bayesian formulation of the image segmentation problem, with a specific focus on the question of the **spatial regularization**. In particular, we introduced a novel prior based on a Gaussian process, together with optimization techniques allowing the method to scale to large images. This new prior was compared to 5 other priors, i.e. the MRF, CRF, TV, FDSP and GLSP priors, within a common Bayesian image segmentation model based on variational inference. The model tractability was achieved by leveraging the concept of local lower bounds on the marginal likelihood. Comparison of the selected spatial priors was performed with respect to several criteria, including the complexity of their inference, their regularization strength, the possibility to estimate automatically all hyperparameters and their impact on uncertainty quantification. Moreover, we showed how our framework may be used to segment specific structures in an image using a narrow band approach. In addition, we introduced an incremental algorithm for the GLSP prior inspired from sparse Bayesian learning allowing the level of spatial regularization to be adapted locally in the image in a data-driven way. To the best of our knowledge, this incremental algorithm is the first attempt to make the level of spatial regularization dependent on the location within the image.

In chapter 4, we proposed a direct application of the methodology developed in the first chapter for **unsupervised quality control of image segmentation**. Our approach is based on a simple probabilistic model that integrates appearance and regularization assumptions in a data-driven way, in contrast to previous unsupervised

methods requiring the user to define a trade-off between the two terms. Our framework does not require any training set or any prior knowledge regarding the segmented structures and provides a result that is visually interpretable. Its effectiveness to extract atypical cases was demonstrated on several large photographic and medical datasets. In particular, it was shown to outperform the classical score-based unsupervised methods.

In chapter 5, we developed a new **robust approach to estimate a consensus from several continuous segmentation maps** produced by different experts or algorithms. The originality of our approach lies in the replacement of the classical Gaussian distribution by heavy-tailed distributions. The possibility to represent these distributions as Gaussian scale mixtures makes our method tractable and, importantly, enables a local assessment of the raters' performances. Therefore, in contrast to the classical Gaussian model, the raters' contributions to the consensus are not uniform in the image but may vary spatially depending on the local performances of the raters. Moreover, we introduced the concept of mixture of consensuses, which is another approach for robust consensus estimation and allows outliers among raters to be identified. In addition, we demonstrated the relevance of the mixture model to cluster the raters over a batch of images.

The last part of the thesis was dedicated to lung cancer screening by LDCT. We proposed a **fully automated pipeline for lung cancer screening based on deep learning** which takes a chest CT scan as input and outputs the associated lung cancer probability. Our pipeline is composed of 3 neural networks corresponding to the 3 following steps: lung segmentation, nodule detection and nodule characterization. The results at the nodule level are then aggregated in a simple manner to produce a prediction at the scan level. Our pipeline was trained on the LIDC-IDRI database, with subjective radiological annotations, and tested on three independent test sets, for which reliable labels were available. Despite the numerous limitations of the training set, our pipeline led to close to state of the art results on the test sets. Moreover, our nodule detector showed a good ability to detect lesions one year before radiologists and we obtained a final specificity close to that of radiologists.

7.2 Perspectives

We discuss in this last section some perspectives for an extension of the approaches developed in the thesis.

7.2.1 Spatial regularization in neural networks

Chapters 4 and 5 are direct examples of application of the methodology developed in chapter 3. A remaining important question is the local adaptivity of the spatial regularisation. We began to explore this issue with the introduction of the incremental algorithm proposed for the GLSP prior. The approach, inspired from sparse Bayesian learning, selects automatically the most relevant basis functions thus adapting locally the level of regularization. The method is nonetheless specific to the GLSP prior and the question of local adaptivity is still open for the other priors.

Moreover, state of the art segmentation performances are now achieved by deep learning-based approaches. Common losses used for image segmentation such as soft Dice or cross entropy do not enforce explicitly spatial regularization. Theoretically,

given unlimited training resources, we could expect the network to learn the appropriate level of spatial regularization automatically. Yet, networks often suffer in practice from limited expressiveness and it is known that, in such cases, the introduction of explicit constraints in the loss function can lead to better results and faster convergence. For instance, [Dong et al., 2020] added a texture term to the classical soft Dice loss, enabling to constrain the appearance along the segmentation boundaries. The proposed loss led to improved performances for gastric antrum segmentation on ultrasound images. In [Clough et al., 2020], the loss is designed to incorporate prior knowledge about the topology of the structures to be segmented, and was demonstrated to achieve better segmentation accuracy on cardiac images.

More generally, there is a rising interest of the computer vision community in moving towards an integration of classical and deep learning-based methods, in particular to improve the interpretability and the uncertainty assessment of deep learning results. A promising avenue for future work would therefore be to explore the generalization of the methodology developed for Bayesian spatial regularization to neural networks. One can note that a related work has recently been proposed by [Liu et al., 2020]. They introduce a variational interpretation of the softmax function, which is typically used after the last layer in neural networks to produce pixel-wise segmentation probabilities. They show how this re-formulation enables to enforce explicitly spatial and shape regularization. The approach was successfully tested on skin lesions segmentation, as shown in Fig. 7.1, which demonstrates again the relevance of incorporating prior knowledge in loss functions.

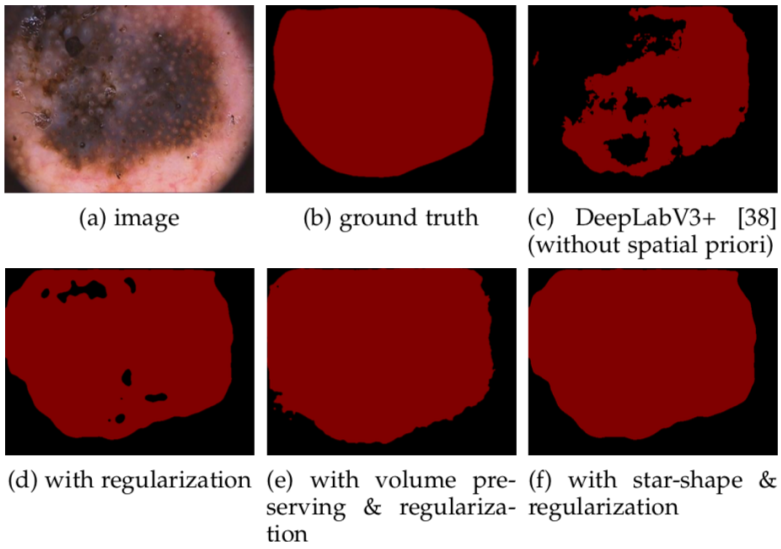


Fig. 7.1: Influence of adding spatial and topological constraints to the training loss on the segmentation results of skin lesions [Liu et al., 2020].

7.2.2 Unsupervised quality control of bounding boxes

A first possibility to extend the framework proposed in chapter 4 for unsupervised quality control of segmentations would be to change the model assumptions. For example, the spatial smoothness hypothesis could be combined with a prior constraining the shape of the segmented structure. This could lead to improved performances when dealing with the segmentation of specific objects. However, the model would loose its genericity and

require a set of trusted segmentations for the configuration of the shape prior parameters. We would then move away from the initial objective, which was to study unsupervised methods.

A second extension possibility would be the correction of segmentations, which is the next step after the detection of suspicious cases. We have shown that two great advantages of our method in comparison to score-based approaches are its interpretability and the possibility to localize potential errors within the image. Our approach could be extended to propose an interactive procedure for the manual delineation of structures in an image, that would automatically assess the quality of the segmentation and suggest alternatives for suspicious segments.

Furthermore, although we restrained the analysis to segmentation labels in this thesis, we believe that our approach could be used on other types of annotations. In particular, it could be of interest for the quality control of bounding box annotations. It is an important issue in the computer vision domain as flawed labels can affect the training and the final performances of deep learning-based detectors. Drawing bounding boxes around objects to be detected is a tedious and time-consuming task. Several approaches have been proposed to alleviate the labelling burden, for instance strategies leveraging crowd-sourcing [Su et al., 2012; Lin et al., 2014], automated approaches [Wu et al., 2020] or new training procedures incorporating a human verification step [Papadopoulos et al., 2016]. Importantly, all require the implementation of a proper quality control, which can be challenging depending on the size of the datasets.

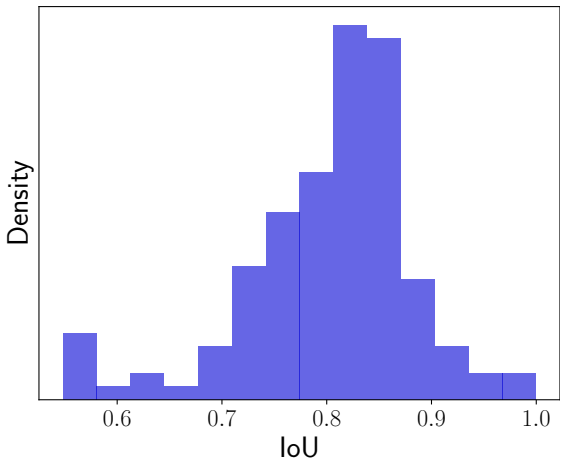


Fig. 7.2: IoU distribution for the analysis of 2D ground truth bounding boxes from the LUNA16 dataset.

Despite its relevance, the question of automatic quality control of bounding box annotations has received little attention. Fig. 7.2 and 7.3 present some preliminary results highlighting the potential of our method for the unsupervised quality control of bounding box labels. The inputs are the nodule bounding box annotations of the LUNA16 challenge [Setio et al., 2017], that was introduced in chapter 6. These annotations include the coordinates of the center of the lesion and its diameter. The approach is tested here in 2D on the middle axial slice of the 127 nodules having a minimum 15 mm diameter. The appearance model of the foreground region is fitted on the image region defined by the input bounding box. The background model is learnt on a tight narrow band surrounding the input bounding box. Our probabilistic model generates a segmentation

used to define a new bounding box, which is compared to the input by computing an intersection over union (IoU) score.

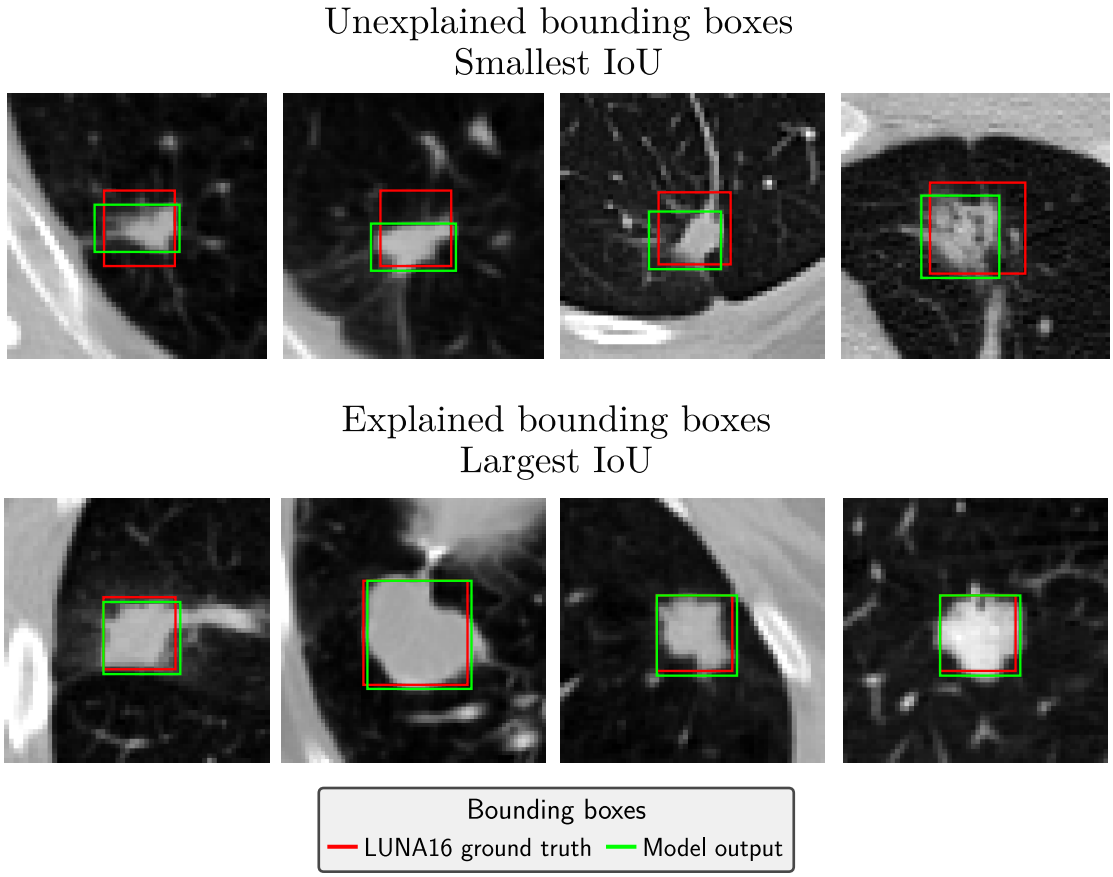


Fig. 7.3: Bounding boxes with the smallest and largest IoU taken from the left and right tails of the distribution, respectively. Cases are ranked according to their IoU value (Largest values to the right).

The analysis is then carried out as for the segmentation labels. The right tail of the histogram in Fig. 7.2 contains cases with the largest IoU scores that are explained by our model. In contrast, left tail samples are characterized by discrepancies between the LUNA16 bounding box and the one produced by our model. As shown in Fig. 7.3, these samples indeed appear to be more suspicious than cases from the right tail. These exploratory results are compelling and demonstrate that the quality control of bounding box annotations is another promising application perspective for our unsupervised framework.

7.2.3 Consensus estimation in a supervised learning setting

Several extensions possibilities for the robust consensus estimation model have already been discussed at the end of chapter 5. The important remark is that our model is not specific to image segmentation and could find an application in other domains where continuous data fusion is required, eventually by adapting the prior defined over the consensus map.

In the thesis, the robust model was mainly explored for the estimation of a consensus between several experts or segmentation algorithms. However, another interesting feature of the framework is its ability to parametrize the raters' performances. This could be used to address a crucial question that arises when developing segmentation algorithms, i.e. how comparable are their performances to those of human experts? More precisely, our model could be applied over a set of images to jointly estimate the performances of the algorithm and of the experts. The next step would then consist in a variance analysis to determine if the average performance of the algorithm lies within the variance of the raters' performances.

Moreover, as for chapter 3, an interesting research perspective would be to study how our framework could be translated to the supervised learning setting, and in particular be used to improve the performances of deep learning-based approaches. Multiple labels can indeed have adverse effects during training and there is no consensus yet regarding the best way to treat them. [Zhang et al., 2020] proposed recently a novel training procedure towards this direction. Inspired by the binary STAPLE algorithm, their framework is composed of 2 networks, shown in Fig. 7.4, that jointly estimate the consensus segmentation and the raters' performances, leading to better results than the commonly used majority voting approach. In contrast to the classical STAPLE algorithm, this deep learning-based approach leverages the whole training set to learn the raters' performances and is thus able to take correlations between images into account.

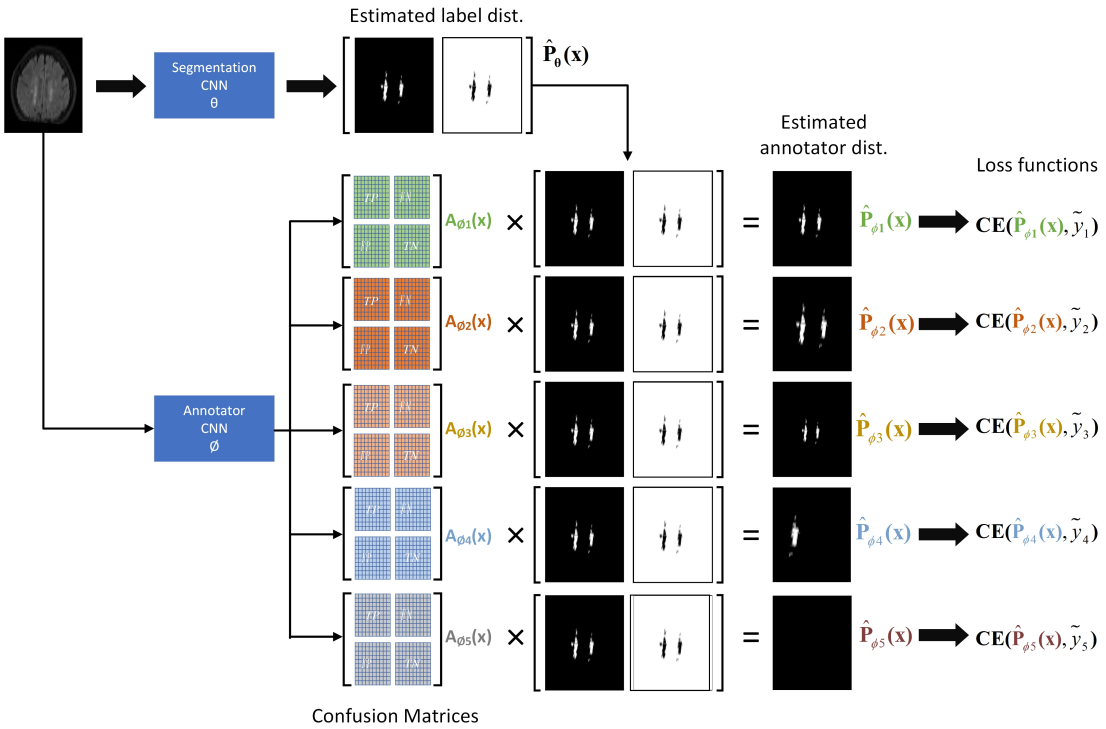


Fig. 7.4: Deep learning framework allowing the performances of annotators and the true segmentation map to be estimated at the same time [Zhang et al., 2020].

7.2.4 Towards an integrative analysis of chest CT scans

Our work on lung cancer screening highlighted the need in datasets with high quality annotations. This obstacle would need to be addressed to enable the improvement of computerized pipelines, but also their fair evaluation reflecting the real performances. The characterization of nodules on LDCT has been demonstrated to be a very difficult task. Combining LDCT with other imaging modalities could be a possibility to identify more specific features associated with malignancy. A monitoring over time would also certainly lead to a finer characterization, as it would allow the growth of the lesion to be taken into account. The latter is indeed a key parameter often associated with malignancy that cannot be assessed with only one time point.

Moreover, as discussed at the end of chapter 6, the future of lung cancer screening certainly lies in the combination of multiple data sources, for instance, imaging resources and biological samples [Benzaquen et al., 2019]. How to integrate such heterogeneous data sources in a same model to predict lung cancer is an open challenge for future work.

Furthermore, the consequences of the current COVID-19 pandemic on lung cancer patients is not fully known. Some results already indicate that the presence of lung cancer is an increased risk factor for developing a severe form of COVID-19 [Luo et al., 2020] but the underlying relationships between the two diseases need to be established, especially to assess the long-term impact of COVID-19 on lung cancer. In addition, the pandemic has created serious disruptions in the lung cancer screening programs [Van Haren et al., 2021], and there are also concerns that COVID-19 might affect the detection of lung cancer based on LDCT. [Calabrò et al., 2020] mentioned indeed that COVID-19 may lead to patterns similar to those found in some lung cancer patients on chest CT scans. The impact of this potential overlap of features between COVID-19 and lung cancer on the performances of computerized screening pipelines will need to be investigated.

Finally, it would be ineffective to reduce the CT scan examination to lung cancer detection, since it may contain information relevant for other lung diseases. In the long term, one could expect a lung cancer screening pipeline to be able not only to detect lung cancer cases but also to notify potential abnormalities pertinent for the general health of the patient.

Appendix A

Spatial priors for Bayesian image segmentation

Contents

A.1 Variational updates . . . 143
A.1.1 Update of $q(Z)$ . . . 143
A.1.2 Update of $q(\mathbf{W})$ . . . 144
A.1.3 Update of α . . . 146
A.1.4 Update of ξ and u . . . 146
A.2 Lower bound . . . 146
A.2.1 Expectations involving appearance parameters . . . 147
A.2.2 Expectations involving the label variable . . . 148
A.2.3 Expectations involving the spatial smoothness variables . . . 149
A.3 RKHS regularizers for the GLSP prior . . . 149
A.4 Derivation of the incremental algorithm for the GLSP prior 151

A.1 Variational updates

In this section, we give the derivations of the variational updates formula. We focus on the label and spatial smoothness variables, since formulas for the appearance parameters are classical solutions of variational Gaussian mixtures (more details can be found for instance in [Bishop, 2006]). The model log joint probability $p(I, Z, \Gamma, \mathbf{W}, \theta_I)$ factorizes as $p(I|Z, \Gamma, \theta_I)p(\Gamma|Z, \theta_I)p(\theta_I)p(Z|\mathbf{W})p(\mathbf{W})$.

A.1.1 Update of $q(Z)$

Eq. 3.19 applied to the label posterior approximation gives:

$$\log q^*(Z) = \mathbb{E}[\log p(I|Z, \Gamma, \theta_I) + \log p(\Gamma|Z, \theta_I) + \log p(Z|\mathbf{W})] + cst\,. \tag{A.1}$$

Developing the first term, we get:

$$\begin{aligned}\mathbb{E}[\log p(I|Z,\Gamma,\mu,\mathbf{\Lambda})] &= \sum_{n=1}^{N}\sum_{k=0}^{1} Z_{nk}\mathbb{E}\left[\sum_{m=1}^{M_k}\Gamma_{nkm}\log\mathcal{N}(I_n;\mu_{km},\mathbf{\Lambda}_{km})\right], \\ &= \sum_{n=1}^{N}\sum_{k=0}^{1} Z_{nk}\sum_{m=1}^{M_k}\delta_{nkm}\left[-\frac{D}{2}\log(2\pi)+\frac{1}{2}\mathbb{E}[\log|\mathbf{\Lambda}_{km}|]\right. \\ &\left.-\frac{1}{2}\mathbb{E}[(I_n-\mu_{km})^T\mathbf{\Lambda}_{km}(I_n-\mu_{km})]\right],\end{aligned} \tag{A.2}$$

with $\mathbb{E}[\log|\mathbf{\Lambda}_{km}|] = \sum_{d=1}^{D}\psi\left(\frac{\nu_{km}+1-i}{2}\right)+D\log 2+\log|W_{km}|$ and $\mathbb{E}[(I_n-\mu_{km})^T\mathbf{\Lambda}_{km}(I_n-\mu_{km})] = D\beta_{km}^{-1}+\nu_{km}(I_n-m_{km})^T W_{km}(I_n-m_{km})$.

The second term leads to:

$$\begin{aligned}\mathbb{E}[\log p(\Gamma|Z,\pi)] &= \sum_{n=1}^{N}\sum_{k=0}^{1} Z_{nk}\mathbb{E}\left[\sum_{m=1}^{M_k}\Gamma_{nkm}\log\pi_{km}\right], \\ &= \sum_{n=1}^{N}\sum_{k=0}^{1} Z_{nk}\sum_{m=1}^{M_k}\delta_{nkm}\mathbb{E}[\log\pi_{km}],\end{aligned} \tag{A.3}$$

with $\mathbb{E}[\log\pi_{km}] = \psi(\gamma_{km})-\psi(\hat{\gamma}_k)$, where $\hat{\gamma}_k = \sum_m \gamma_{km}$.

The last term involves the label prior which leads to:

$$\begin{aligned}\mathbb{E}[\log p(Z|\mathbf{W})] &= \sum_{n=1}^{N}\mathbb{E}\left[\log\left(\left(\frac{1}{1+\exp y_n}\right)^{1-Z_n}\left(\frac{\exp y_n}{1+\exp y_n}\right)^{Z_n}\right)\right], \\ &= Z_n\mathbb{E}\left[y_n\right].\end{aligned} \tag{A.4}$$

where y_n is equal to w_n for the TV, FDSP and GP priors, and $y_n = \mathbf{\Phi}_n^T\mathbf{W}$ for the GLSP prior.

Summing the three expectations A.2, A.3 and A.4 and taking the exponential, we get $q^*(Z) \propto \prod_n \rho_{n0}^{1-Z_n}\rho_{n1}^{Z_n}$, where the expressions for ρ_{n0} and ρ_{n1} are given by Eq. 3.20. We finally obtain a product of Bernoulli distributions with parameters η_{n0} and η_{n1}, such that $q^*(Z) = \prod_n \eta_{n0}^{1-Z_n}\eta_{n1}^{Z_n}$.

A.1.2 Update of $q(\mathbf{W})$

Eq. 3.19 applied to the posterior approximation of the weights gives:

$$\begin{aligned}\log q^*(\mathbf{W}) &= \mathbb{E}[\log p(Z|\mathbf{W})+\log p(\mathbf{W})]+cst\,, \\ &\geqslant \mathbb{E}[\log F(Z,\mathbf{W},\xi)+\log p(\mathbf{W})]+cst\,.\end{aligned} \tag{A.5}$$

JJ bound

For the GLSP, TV and FDSP priors, F is obtained through the JJ bound. Discarding the terms independent with respect to $\mathbf{W}$, $\log F(Z, \mathbf{W}, \xi)$ is written:

$$\log F(Z, \mathbf{W}, \xi) = \sum_{n=1}^{N} -\lambda(\xi_n) y_n^2 + y_n \left(\eta_{n1} - \frac{1}{2}\right) + cst\,, \tag{A.6}$$

where cst is a constant independent of $\mathbf{W}$, and with $y_n = w_n$ for the TV and FDSP priors or $y_n = \mathbf{\Phi}_n^T \mathbf{W}$ for the GLSP prior.

The second term in Eq. A.5 depends on the spatial prior. For the GLSP, $\mathbb{E}[\log p(\mathbf{W})] = -\frac{\alpha}{2}\mathbf{W}^T\mathbf{R}\mathbf{W}$, which leads to:

$$\log q^*(\mathbf{W}) = -\frac{1}{2}\mathbf{W}^T \left[\mathbf{\Phi}\mathbf{B}\mathbf{\Phi}^T + \alpha\mathbf{R}\right]\mathbf{W} + \sum_{n=1}^{N}\left(\eta_{n1} - \frac{1}{2}\right)\mathbf{\Phi}_n^T\mathbf{W} + cst\,. \tag{A.7}$$

By identifying the quadratic and linear terms in $\mathbf{W}$, we recognize a Gaussian distribution of parameters given by Eq. 3.31 and Eq. 3.32.

For the TV and FDSP priors, we further assume a factorization between the $q(w_n)$. Keeping the terms involving w_n, we get for the TV prior:

$$\begin{aligned}\log q^*(w_n) = &-\frac{1}{2}w_n^2\left(2\lambda(\xi_n) + \sum_d \frac{\alpha}{\sqrt{u_n}} + \sum_d \frac{\alpha}{\sqrt{u_{\xi_{d(n)}}}}\right)\\ &+ w_n\left(\eta_{n1} - \frac{1}{2} + \alpha\sum_d \frac{\mathbb{E}[w_{\delta_{d(n)}}]}{\sqrt{u_n}} + \alpha\sum_d \frac{\mathbb{E}[w_{\xi_{d(n)}}]}{\sqrt{u_{\xi_{d(n)}}}}\right) + cst\,.\end{aligned} \tag{A.8}$$

We recognize again a Gaussian distribution whose parameters are given by Eq. 3.33 and Eq. 3.34. Following the same approach for the FDSP prior, we have:

$$\begin{aligned}\log q^*(w_n) = &-\frac{1}{2}w_n^2\left(2\lambda(\xi_n) + \sum_d \frac{\alpha}{2} + \sum_d \frac{\alpha}{2}\right)\\ &+ w_n\left(\eta_{n1} - \frac{1}{2} + \frac{\alpha}{2}\sum_d \mathbb{E}[w_{\delta_d(n+2)}] + \frac{\alpha}{2}\sum_d \mathbb{E}[w_{\delta_d(n-2)}]\right) + cst\,.\end{aligned} \tag{A.9}$$

Thus $q^*(\mathbf{W})$ is also a Gaussian distribution whose parameters are given by Eq. 3.35 and Eq. 3.36.

Böhning bound

Following the same approach, we can write for the Böhning bound:

$$\log F(Z, \mathbf{W}, \xi) = \sum_{n=1}^{N} -\frac{1}{2}aw_n^2 + w_n(b_n + \eta_{n1}) + cst\,. \tag{A.10}$$

Adding the prior term, we finally obtain:

$$q^*(\mathbf{W}) = -\frac{1}{2}\mathbf{W}^T \left(a\mathbf{I}_N + \mathbf{\Sigma}_{\text{GP}}^{-1}\right) \mathbf{W} + \mathbf{W}[\eta_1 + B] + cst\,, \tag{A.11}$$

which leads to Eq. 3.37 and Eq. 3.38.

A.1.3 Update of α

α is the hyperparameter of the TV, FDSP and GLSP priors and controls the strength of the regularization. It can be estimated automatically in a data-driven way. Considering α as a variable, we note $q(\alpha)$ the approximation of its posterior distribution, assumed to be a Dirac distribution. The mode of $q(\alpha)$ is found by maximizing Eq. 3.19, which leads to the update formula Eq. 3.39, Eq. 3.40 and Eq. 3.41.

A.1.4 Update of ξ and u

ξ is the additional variational parameter introduced by the local variational bounds. We follow the same approach as for α and assume $q(\xi_n)$ to be a Dirac distribution. Eq. 3.19 applied to $q(\xi_n)$ gives for the JJ bound:

$$\log q^*(\xi_n) = \log \sigma(\xi_n) - \frac{\xi_n}{2} - \lambda(\xi_n)(\mathbb{E}[y_n^2] - \xi_n^2)\,. \tag{A.12}$$

Taking the derivative with respect to ξ_n and setting it to zero gives $\lambda'(\xi_n)(\mathbb{E}[y_n^2] - \xi_n^2) = 0$. As $\lambda'(\xi_n) \leq 0$, we obtained the formula reported in section 3.4.3.

Similarly, we have for the Böhning bound:

$$\log q^*(\xi) = [\frac{1}{4}\mathbf{I}_N\xi - g(\xi)]^T\mathbb{E}[\mathbf{W}] - \frac{1}{8}\xi^T\xi + g(\xi)^T\xi - \text{lse}(\xi)\,. \tag{A.13}$$

Zeroing the derivatives leads to the update of ξ for the Böhning bound reported in section 3.4.3.

Finally, u is introduced by the bound over the square root function that allows the TV prior to be tractable. Taking the derivative of Eq. 3.19 applied to $q(u_n)$ gives:

$$\frac{u_n - \sum_{d=1}^{D}\mathbb{E}[(w_n - w_{\delta_{d(n)}})^2]}{2u_n\sqrt{u_n}} = 0\,. \tag{A.14}$$

It therefore leads to the update presented in section 3.4.3.

A.2 Lower bound

In this paper, variational inference is used to estimate the model parameters and to maximize a lower bound $\mathcal{L}(q)$ over the data log likelihood. Computing the lower bound is interesting for several reasons. First, it allows the convergence of the model to be assessed in a convenient manner. Second, it provides a way to implement a quality check as each iteration of the model should correspond to an increase in the lower bound.

Finally, the lower bound is also a useful tool to perform model selection by comparing the values reached after convergence. However, performing model selection requires to be able to compute the lower bound with the constants included, while the latter may be neglected for a simple convergence monitoring. In particular, this is not possible for the TV, MRF and CRF priors due to the intractability of their normalization constants. In this case, the lower bound can only be computed up to a constant, which prevents any application to model selection.

For the MRF and CRF priors, the lower bound can be rewritten as:

$$\begin{aligned}\mathcal{L}(q) = \mathbb{E}[\log p(I|\Gamma, Z, \theta_I)] + \mathbb{E}[\log p(\Gamma|Z, \theta_I)] + \mathbb{E}[\log p(\theta_I)] + \mathbb{E}[\log P(Z)] \\ - \mathbb{E}[\log q(Z)] - \mathbb{E}[\log q(\Gamma)] - \mathbb{E}[q(\theta_I)]\,.\end{aligned} \tag{A.15}$$

For the other priors, Eq. 3.30 leads to:

$$\begin{aligned}\mathcal{J}(q) = \mathbb{E}[\log p(I|\Gamma, Z, \theta_I)] + \mathbb{E}[\log p(\Gamma|Z, \theta_I)] + \mathbb{E}[\log p(\theta_I)] \\ + \mathbb{E}[\log F(Z, \mathbf{W}, \xi)] + \mathbb{E}[\log p(\mathbf{W})] \\ - \mathbb{E}[\log q(Z)] - \mathbb{E}[\log q(\Gamma)] - \mathbb{E}[q(\mathbf{W})] - \mathbb{E}[q(\theta_I)]\,.\end{aligned} \tag{A.16}$$

We recall that $\theta_I = \{\mu, \mathbf{\Lambda}, \pi\}$ represents the intensity variables. Therefore, $\mathbb{E}[\log p(\theta_I)]$ and $\mathbb{E}[\log q(\theta_I)]$ can be expanded as $\mathbb{E}[\log p(\mu|\mathbf{\Lambda})] + \mathbb{E}[\log p(\mathbf{\Lambda})] + \mathbb{E}[\log p(\pi)]$ and $\mathbb{E}[\log q(\mu|\mathbf{\Lambda})] + \mathbb{E}[\log q(\mathbf{\Lambda})] + \mathbb{E}[\log q(\pi)]$, respectively.

The values of the different expectations are reported in the following sections.

A.2.1 Expectations involving appearance parameters

Formula given in this section are classical results for variational mixtures of Gaussian distributions. More details can be found in [Bishop, 2006].

$$\begin{aligned}\mathbb{E}[\log p(I|Z, \Gamma, \mu, \mathbf{\Lambda})] = \frac{1}{2}\sum_{k=0}^{1}\sum_{m=1}^{M_k} N_{km}\Big[\mathbb{E}[\log|\mathbf{\Lambda}_{km}|] - \frac{D}{\beta_{km}} - \nu_{km}\operatorname{Tr}(S_{km}W_{km}) \\ - \nu_{km}(\overline{I}_{km} - m_{km})^T W_{km}(\overline{I}_{km} - m_{km}) - D\log(2\pi)\Big]\,.\end{aligned} \tag{A.17}$$

$$\mathbb{E}[\log p(\Gamma|Z, \pi)] = \sum_{n=1}^{N}\sum_{k=0}^{1}\sum_{m=1}^{M_k}\eta_{nk}\delta_{nkm}\mathbb{E}[\log\pi_{km}]\,. \tag{A.18}$$

$$\mathbb{E}[\log p(\pi)] = \sum_{k=0}^{1}\left\{\log C(\gamma_{k0}) + (\gamma_{k0} - 1)\sum_{m=1}^{M_k}\mathbb{E}[\log\pi_{km}]\right\}\,. \tag{A.19}$$

$$\mathbb{E}[\log p(\mu,\mathbf{\Lambda})] = \sum_{k=0}^{1}\left\{\frac{1}{2}\sum_{m=1}^{M_k}\Big[D\log\frac{\beta_{k0}}{2\pi}+\mathbb{E}[\log|\mathbf{\Lambda}_{km}|]-\frac{D\beta_{k0}}{\beta_{km}} - \beta_{k0}\nu_{km}(m_{km}-m_{k0})^T W_{km}(m_{km}-m_{k0})\Big] + M_k\log B(W_{k0},\nu_{k0}) + \frac{\nu_{k0}-D-1}{2}\sum_{m=1}^{M_k}\mathbb{E}[\log|\mathbf{\Lambda}_{km}|] - \frac{1}{2}\sum_{m=1}^{M_k}\nu_{km}\operatorname{Tr}(W_{k0}^{-1}W_{km})\right\}, \tag{A.20}$$

where $B(W,\nu) = |W|^{-\nu/2}\left(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}$.

The entropic terms are given by:

$$\mathbb{E}[\log q(\Gamma)] = \sum_{n=1}^{N}\sum_{k=0}^{1}\sum_{m=1}^{M_k}\delta_{nkm}\log\delta_{nkm}\,. \tag{A.21}$$

$$\mathbb{E}[\log q(\pi)] = \sum_{k=0}^{1}\left\{\sum_{m=1}^{M_k}((\gamma_{km}-1)\mathbb{E}[\log\pi_{km}]) + \log C(\gamma_k)\right\}, \tag{A.22}$$

where $C(\gamma_k) = \Gamma(\sum_m\gamma_{km})/\prod_m\Gamma(\gamma_{km})$.

$$\mathbb{E}[\log q(\mu,\mathbf{\Lambda})] = \sum_{k=0}^{1}\sum_{m=1}^{M_k}\frac{1}{2}\mathbb{E}[\log|\mathbf{\Lambda}_{km}|] + \frac{D}{2}\log\left(\frac{\beta_{km}}{2\pi}\right) - \frac{D}{2} - \mathrm{H}[q(\mathbf{\Lambda}_{km})]\,, \tag{A.23}$$

where $H[\mathbf{\Lambda}] = -\log B(W,\nu) - \frac{\nu-D-1}{2}\mathbb{E}[\log|\mathbf{\Lambda}|] + \frac{\nu D}{2}$.

A.2.2 Expectations involving the label variable

The expectation of the label prior $\mathbb{E}[\log p(Z)]$ leads to $\frac{\beta}{2}\sum_{n=1}^{N}\sum_{i\in\delta(n)}\sum_{k=0}^{1}\eta_{nk}\eta_{ik} + cst$ and $\frac{\beta}{2}\sum_{n=1}^{N}\sum_{i\in\delta(n)}\sum_{k=0}^{1}\eta_{nk}\eta_{ik}\frac{\exp\left(-\gamma(I_n-I_i)^2\right)}{\mathrm{dist}(n,i)} + cst$ for the MRF and CRF priors, respectively. The constants cannot be computed because of the intractability of the partition functions of the priors.

For the other priors, the expectation of the label prior $\mathbb{E}[\log p(Z|\mathbf{W})]$ is replaced by a local variational bound $\mathbb{E}[\log F(Z,\mathbf{W},\xi)]$ to allow tractability. For the JJ bound, we have:

$$\mathbb{E}[\log F(Z,\mathbf{W},\xi)] = \sum_{n=1}^{N}\mathbb{E}[y_n]\left(\eta_{n1}-\frac{1}{2}\right) + \log\sigma(\xi_n) - \frac{\xi_n}{2} - \lambda(\xi_n)(\mathbb{E}[y_n^2]-\xi_n^2)\,, \tag{A.24}$$

where $y_n = w_n$ for the TV and FDSP priors, and $y_n = \mathbf{\Phi}_n^T \mathbf{W}$ for the GLSP prior. We use the Böhning bound for the GP prior which leads to:

$$\mathbb{E}[\log F(Z, \mathbf{W}, \xi)] = -\frac{1}{2} a \operatorname{Tr}(\mathbf{\Sigma}_{\mathbf{W}}) + \sum_{n=1}^{N} \eta_{n1} \mu_{\mathbf{W}_n} - \frac{1}{2} a \mu_{\mathbf{W}_n}^2 + b_n \mu_{\mathbf{W}_n} - c_n \,, \quad \text{(A.25)}$$

with $a = 1/4$, $b_n = a\xi_n - g(\xi_n)$ and $c_n = a\xi_n^2/2 - g(\xi_N)\xi_n + \operatorname{lse}(\xi_n)$. We recall that g is the gradient of the LogSumExp function.

The entropic term is always given by:

$$\mathbb{E}[\log q(Z)] = \sum_{n=1}^{N} \eta_{n1} \log \eta_{n1} + (1 - \eta_{n1}) \log(1 - \eta_{n1}) \,. \quad \text{(A.26)}$$

A.2.3 Expectations involving the spatial smoothness variables

The expectation of the prior over $\mathbf{W}$ depends on the chosen spatial regularization. For a TV prior, the expectation can only be computed up to an additive constant because of the intractability of the normalization constant, which leads to $\mathbb{E}[\log p(\mathbf{W})] = N(\log \alpha - 1) + cst$. In contrast, we have $\mathbb{E}[\log p(\mathbf{W})] = \frac{1}{2}\left(-N \log 4\pi + N \log \alpha + \log |\mathbf{\Lambda}_{\text{FDSP}}| - N\right)$ for a FDSP prior and $\mathbb{E}[\log p(\mathbf{W})] = \frac{1}{2}(L \log \alpha + \log |\mathbf{R}| - L \log 2\pi - L)$ for a GLSP prior. Finally, with a GP prior $\mathcal{N}(\mathbf{W}; 0, \mathbf{\Sigma}_{\text{GP}})$ we obtain:

$$\mathbb{E}[\log p(\mathbf{W})] = -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{\Sigma}_{\text{GP}}| - \frac{1}{2} \left[\operatorname{Tr}\left(\mathbf{\Sigma}_{\text{GP}}^{-1} \mathbf{\Sigma}_{\mathbf{W}} \right) + \mu_{\mathbf{W}}^T \mathbf{\Sigma}_{\text{GP}}^{-1} \mu_{\mathbf{W}} \right] . \quad \text{(A.27)}$$

The entropic term is written $\mathbb{E}[\log q(\mathbf{W})] = -\frac{1}{2}\left(L \log 2\pi + \log |\mathbf{\Sigma}_{\mathbf{W}}| + L\right)$ and $\mathbb{E}[\log q(\mathbf{W})] = -\frac{1}{2}\left(N \log 2\pi + \log |\mathbf{\Sigma}_{\mathbf{W}}| + N\right)$ for the GP and GLSP priors, respectively. For the TV and FDSP priors, it is $\mathbb{E}[\log q(\mathbf{W})] = -\frac{1}{2}(N \log 2\pi + N + \sum_{n=1}^{N} \log |\mathbf{\Sigma}_{w_n}|)$, because we further assume that $q(\mathbf{W}) = \prod_n q(w_n)$.

A.3 RKHS regularizers for the GLSP prior

In this section, we show how to obtain regularizers of the label field based on RKHS (reproducing kernel Hilbert space) methods. We recall that the label field $f : \mathbb{R}^D \longmapsto \mathbb{R}$ is given by:

$$\forall x \in \mathbb{R}^D, \; l(x) = \sum_{l=1}^{L} \Phi_l(x) w_l \,. \quad \text{(A.28)}$$

Let S be a $D \times D$ symmetric positive definite matrix, $K_S(x, y) = \exp -\frac{1}{2}(x - y)^T S^{-1}(x - y)$ a translation invariant kernel and $\widehat{K_S} = \int_{\mathbb{R}^D} e^{-ix\xi} K_S(x) dx$ its Fourier transform. The space $\mathcal{H}_{K_S}$ consisting of integrable d-vector fields $g : \mathbb{R}^D \longrightarrow \mathbb{R}$ such that:

$$||g||_{K_S}^2 = \frac{1}{(2\pi)^D} \int_{\mathbb{R}^D} \frac{|\widehat{g}(\xi)|^2}{\widehat{K_S}(\xi)} d\xi < +\infty \,, \quad \text{(A.29)}$$

endowed with the inner product:

$$\langle g, h\rangle_{\mathcal{H}_{K_S}} = \frac{1}{(2\pi)^D}\int_{\mathbb{R}^D}\frac{\widehat{g}(\xi)\widehat{h}(\xi)^*}{\widehat{K_S}(\xi)}d\xi\,, \tag{A.30}$$

is an RKHS with K_S as the reproducing kernel. Due to the reproducing property, the following inequality holds for any $g \in \mathcal{H}_{K_S}$ and any $x, x' \in \mathbb{R}^D$:

$$|g(x) - g(x')| \leq ||g||_{\mathcal{H}_{K_S}}||K_S(x) - K_S(x')||_{\mathcal{H}_{K_S}}\,. \tag{A.31}$$

The RKHS norm, $||.||_{\mathcal{H}_{K_S}}$, thus controls the variations of the function with respect to the geometry induced by the kernel, K_S. Moreover, the partial derivatives of $f \in \mathcal{H}_{K_S}$ exist and all lie in $\mathcal{H}_K$ [Zhou, 2008].

The label field f is a linear combination of Gaussian basis functions Φ_k that are in $\mathcal{H}_{K_S}$. Therefore, penalizing the RKHS norm of the label field or of its derivatives encourages the function to present smooth variations over the image, which is the desired output. Let D be a linear differential operator and j a non-negative integer. We can then write:

$$\begin{aligned}\langle D^j f, D^j f\rangle_{\mathcal{H}_{K_S}} &= \langle\sum_l w_l D^j\Phi_l, \sum_k w_k D^j\Phi_k\rangle_{\mathcal{H}_{K_S}}\,,\\ &= \mathbf{W}^T\mathbf{R}\mathbf{W}\,,\end{aligned} \tag{A.32}$$

where $\mathbf{R}$ is an $L \times L$ matrix whose k, l coefficient is $\langle D^j\Phi_k, D^j\Phi_l\rangle_{\mathcal{H}_{K_S}}$ and can be computed in closed form.

From Eq. A.30 and recalling that $\Phi_k(x) = \Phi(x - x_k)$, we can write:

$$\begin{aligned}R_{kl} &= \frac{1}{(2\pi)^D}\int_{\mathbb{R}^D}\widehat{D^j\Phi_k}(\xi)\widehat{D^j\Phi_l}(\xi)^*\widehat{K_S}^{-1}(\xi)d\xi\,,\\ &= \mathcal{F}^{-1}\left[\widehat{D^j\Phi}(\xi)\widehat{D^j\Phi}^*(\xi)\widehat{K_S}^{-1}(\xi)\right](x_k - x_l)\,,\end{aligned} \tag{A.33}$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform.

In the chapter, we consider the cases where $D = \nabla^s$, where $s = 0, 1, 2$ or 3. Recall that for any multi-index $\alpha = (\alpha_1, \ldots, \alpha_D)$, $\mathcal{F}\left[\left(\frac{\partial}{\partial x}\right)^\alpha g\right](x) = (i\xi)^\alpha\hat{g}(\xi) = i^{|\alpha|}\xi_1^{\alpha_1}\xi_2^{\alpha_2}\cdots\xi_D^{\alpha_D}\hat{g}(\xi)$, where $|\alpha| = \sum_k \alpha_k$. Thus for any $g \in \mathcal{H}_{K_S}$, we can write:

$$\begin{aligned}\mathcal{F}\left[\nabla g\right] &= i\xi\hat{g}(\xi)\,,\\ \mathcal{F}\left[\nabla^2 g\right] &= -||\xi||^2\hat{g}(\xi)\,,\\ \mathcal{F}\left[\nabla^3 g\right] &= -i||\xi||^2\xi\hat{g}(\xi)\,.\end{aligned} \tag{A.34}$$

Furthermore, the Fourier transform of a normalized Gaussian basis such that it takes the value 1 at the origin $\Phi_k(x) = \exp -\frac{1}{2}(x - x_k)^T S_k^{-1}(x - x_k)$ is given by $\mathcal{F}[\Phi_k](x) = |2\pi S_k|^{1/2}\exp(-i\xi^T x_k)\exp\left(-\frac{1}{2}\xi^T S_k\xi\right)$. Combining this expression together

with Eq. A.34 in Eq. A.33, we derive a closed-form expression for the k, l coefficient of the matrix $\mathbf{R}$ for normalized Gaussian basis for $s = 0, 1, 2$ or 3:

$$\mathbf{R}_{kl} = \left| \frac{S_k S_l}{S(S_k + S_l - S)} \right|^{1/2} (-\Delta)^s \, K_{S_k + S_l - S}(x_k - x_l) \,. \tag{A.35}$$

The above expression involves powers of the Laplacian of a Gaussian kernel, that can be computed to give:

$$\Delta K_S(x) = \left\{ x^T S^{-2} x - \mathrm{Tr}(S^{-1}) \right\} K_S(x) \,, \tag{A.36}$$

$$\Delta^2 K_S(x) = \left\{ \left(\mathrm{Tr}(S^{-1}) - x^T S^{-2} x \right)^2 + 2\,\mathrm{Tr}(S^{-2}) - 4 x^T S^{-3} x \right\} K_S(x) \,, \tag{A.37}$$

$$\begin{aligned} \Delta^3 K_S(x) = \Big\{ \left(\mathrm{Tr}(S^{-1}) - x^T S^{-2} x \right) \Big[- \left(\mathrm{Tr}(S^{-1}) - x^T S^{-2} x \right)^2 - 6\,\mathrm{Tr}(S^{-2}) \\ + 12 x^T D^{-3} x \Big] + 24 x^T S^{-4} x - 8\,\mathrm{Tr}(S^{-3}) \Big\} K_S(x) \,. \end{aligned} \tag{A.38}$$

A.4 Derivation of the incremental algorithm for the GLSP prior

The incremental algorithm [Tipping & Faul, 2003] is based on a second order Taylor expansion of the lower bound $\mathcal{L}(q) + \log p(\mathbf{W}|\alpha)$. Keeping only the terms that depend on $\mathbf{W}$, gives:

$$\begin{aligned} \mathcal{L}(q) + \log p(\mathbf{W}|\alpha) &= \sum_{n=1}^{N} \sum_{Z_n} q(Z_n) \log p(Z|\mathbf{W}) + \log p(\mathbf{W}|\alpha) + cst \,, \\ &= \sum_{n=1}^{N} \eta_{n1} \log \sigma(\mathbf{\Phi}_n^T \mathbf{W}) + (1 - \eta_{n1}) \log \sigma(-\mathbf{\Phi}_n^T \mathbf{W}) - \frac{1}{2} \mathbf{W}^T \mathbf{A} \mathbf{W} + cst \,. \end{aligned} \tag{A.39}$$

The function optimized by the RVM when the likelihood is a Bernoulli distribution is exactly the same, except that the binary classification targets $t_n \in \{0, 1\}$ are replaced here by the "soft" labels $\eta_{n1} \in [0, 1]$.

Let g_n be the function defined as $g_n(x) = \eta_{n1} \log \sigma(x) + (1 - \eta_{n1}) \log \sigma(-x)$. Eq. A.39 can then be rewritten as:

$$\mathcal{L}(q) + \log p(\mathbf{W}|\alpha) = \sum_{n=1}^{N} g_n(\mathbf{\Phi}_n^T \mathbf{W}) - \frac{1}{2} \mathbf{W}^T \mathbf{A} \mathbf{W} + cst \,. \tag{A.40}$$

Furthermore, we have $g_n'(x) = (\eta_{n1} - \sigma(x))$ and $g_n''(x) = (\sigma(x) - 1)\sigma(x)$. Therefore, differentiating Eq. A.40 twice gives:

$$\frac{\partial(\mathcal{L}(q) + \log p(\mathbf{W}|\alpha))}{\partial \mathbf{W}} = \sum_{n=1}^{N} \mathbf{\Phi}_n g_n'(\mathbf{\Phi}_n^T \mathbf{W}) - \mathbf{A}\mathbf{W} = \mathbf{\Phi} g' - \mathbf{A}\mathbf{W}\,, \tag{A.41}$$

$$\frac{\partial^2(\mathcal{L}(q) + \log p(\mathbf{W}|\alpha))}{\partial \mathbf{W}^2} = \sum_{n=1}^{N} \mathbf{\Phi}_n g_n''(\mathbf{\Phi}_n^T \mathbf{W})\mathbf{\Phi}_n^T - \mathbf{A} = -[\mathbf{\Phi}\mathbf{B}\mathbf{\Phi}^T + \mathbf{A}]\,. \tag{A.42}$$

This leads to the covariance of the Gaussian approximation of the weights posterior reported in Eq. 3.45. In addition, the gradient (Eq. A.41) equals zero at the mode, leading to $\mathbf{\Phi} g' - \mathbf{A}\mu_{\mathbf{W}} = 0$. Introducing the covariance and rearranging, we can finally write the mode as $\mu_{\mathbf{W}} = \mathbf{\Sigma}_{\mathbf{W}}\mathbf{\Phi}\mathbf{B}\hat{\mathbf{t}}$ with $\hat{\mathbf{t}} = \mathbf{B}^{-1}g' + \mathbf{\Phi}^T\mu_{\mathbf{W}}$.

The α_l are updated by maximizing $\mathcal{L}(\alpha)$ obtained after marginalizing out the weight variable from Eq. A.40, the integral being tractable with the Laplace approximation:

$$\begin{aligned}\mathcal{L}(\alpha) &= \int_{\mathbf{W}} \mathcal{L}(q) + \log p(\mathbf{W}|\alpha) d\mathbf{W}\,, \\ &\approx \mathcal{L}(q, \mu_{\mathbf{W}}) + \log p(\mu_{\mathbf{W}}|\alpha) + \frac{L}{2}\log 2\pi + \frac{1}{2}\log|\mathbf{\Sigma}_{\mathbf{W}}|\,.\end{aligned} \tag{A.43}$$

Introducing the matrix $\mathbf{C} = \mathbf{B}^{-1} + \mathbf{\Phi}^T\mathbf{A}\mathbf{\Phi}$, one can show that $\mathcal{L}(\alpha)$ and $\mathcal{F}(\alpha) = -\frac{1}{2}\left\{\log 2\pi + \log|\mathbf{C}| + \hat{\mathbf{t}}^T\mathbf{C}^{-1}\hat{\mathbf{t}}\right\}$ have the same derivative with respect to α_l. The second expression is convenient to estimate the contribution of a particular basis l to the lower bound. All the expressions presented in [Tipping & Faul, 2003] can be used at each iteration to select the basis function with the largest gain in lower bound and the appropriate action to apply (re-estimation, addition or deletion). We refer to the original paper for more details.

Appendix B

Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model

Contents

B.1 Unsupervised indices . . . 153
B.1.1 *Zeb* [Zhang et al., 2008] . . . 153
B.1.2 F_{RC} [Zhang et al., 2008] . . . 154
B.1.3 η [Zhang et al., 2008] . . . 155
B.1.4 *GS* [Johnson & Xie, 2011] . . . 155
B.2 FDSP prior - variational inference . . . 155
B.2.1 Update of $q_Z^*(Z)$ . . . 156
B.2.2 Update of $q_{\mathbf{W}}^*(\mathbf{W})$ . . . 156
B.2.3 Update of $q_\alpha^*(\alpha)$ . . . 157
B.2.4 Update of $q_{\xi_n}^*(\xi_n)$ . . . 157
B.3 FDSP prior - lower bound . . . 157

B.1 Unsupervised indices

We give in this section the formula used to compute the unsupervised indices. We denote by R the number of regions inside an image (typically 2 here, for the foreground and background regions). R_j denotes the set of voxels in region j and $|R_j|$ is the number of voxels in region j. Each indicator requires the computation of an intra-region uniformity metric IU and an inter-region disparity metric ID.

B.1.1 *Zeb* [Zhang et al., 2008]

$$\mathrm{IU}_j = \frac{1}{|R_j|} \sum_{s \in R_j} \max \left\{ \mathrm{contrast}(s,t), t \in W(s) \cap R_j \right\} , \tag{B.1}$$

where $W(s)$ is the neighborhood of voxel s and:

$$\text{contrast}(s,t) = \frac{1}{\nu}\sum_{i=1}^{\nu}\left|I_s^i - I_t^i\right| . \tag{B.2}$$

$$\text{ID}_j = \frac{1}{|b(R_j)|}\sum_{s\in b(R_j)} \max\left\{\text{constrast}(s,t), t\in W(s), t\notin R_j\right\} , \tag{B.3}$$

where $b(R_j)$ is the set of pixels on the border of R_j.

The final indicator is given by:

$$Zeb = \frac{\text{IU}}{\text{ID}} = \frac{\sum_j \text{IU}_j}{\sum_j \text{ID}_j} . \tag{B.4}$$

B.1.2 F_{RC} [Zhang et al., 2008]

$$\text{IU} = \frac{1}{R}\sum_{j=1}^{R}\frac{|R_j|}{N}e^2(R_j) , \tag{B.5}$$

where:

$$e^2(R_j) = \frac{1}{\nu}\sum_{i=1}^{\nu}\sum_{s\in R_j}\left(I_s^i - \hat{I}_{R_j}^i\right)^2 . \tag{B.6}$$

$\hat{I}_{R_j}^i$ is defined for $1\leq i\leq \nu$ by:

$$\hat{I}_{R_j}^i = \frac{1}{|R_j|}\sum_{s\in R_j} I_s^i . \tag{B.7}$$

$$\text{ID} = \frac{1}{R}\sum_{j=1}^{R}\frac{|R_j|}{N}\left(\frac{1}{|W(R_j)|}\sum_{t\in W(R_j)} D(R_j,R_t)\right) , \tag{B.8}$$

where $W(R_j)$ is the set of neighboring regions of R_j and:

$$D(R_j,R_t) = \frac{1}{\nu}\sum_i |\hat{I}_{R_j}^i - \hat{I}_{R_t}^i| . \tag{B.9}$$

The final indicator is given by:

$$F_{\text{RC}} = \text{IU} - \text{ID} . \tag{B.10}$$

B.1.3 η [Zhang et al., 2008]

The background is denoted here by b, while f denotes the foreground.

$$\mathrm{IU} = \frac{N_b}{N} e^2(R_b) + \frac{N_f}{N} e^2(R_f)\,, \tag{B.11}$$

where N_b and N_f are the number of voxels in the background and foreground, respectively, and $e^2(R_j)$ is defined as previously.

$$\mathrm{ID} = \frac{N_b N_f}{N^2} \left(\hat{I}_{R_f} - \hat{I}_{R_b}\right)^2\,, \tag{B.12}$$

where $\hat{I}_{R_j} = \frac{1}{\nu}\sum_{i=1}^{\nu} \hat{I}^i_{R_j}$.

The final indicator is given by:

$$\eta = \frac{\mathrm{IU}}{\mathrm{ID}}\,. \tag{B.13}$$

B.1.4 *GS* [Johnson & Xie, 2011]

$$\mathrm{IU} = \frac{\sum_j |R_j| V_j}{\sum_j |R_j|}\,, \tag{B.14}$$

where V_j is the variance of region j.

The inter-region disparity metric used is the Global Moran's I, defined as:

$$\mathrm{ID} = \frac{R}{\sum_i \sum_{j\neq i} w_{ij}} \frac{\sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{R}(y_i - \overline{y})^2}\,, \tag{B.15}$$

where $w_{ii} = 0$, $w_{ij} = 1$ if R_i and R_j are neighbors and 0 otherwise. y_i is the mean intensity value of region R_i and $\overline{y}$ is the mean intensity value of the image.

The final indicator is given by:

$$\mathrm{GS} = \mathrm{IU} + \mathrm{ID}\,. \tag{B.16}$$

B.2 FDSP prior - variational inference

We present in this section the derivation of the variational update formula (4.11), (4.12), (4.13), (4.14) and (4.15). The likelihood of the model $p(I, Z, \mathbf{W}, \alpha)$ factorizes as $p(I|Z)p(Z|\mathbf{W})p(\mathbf{W}|\alpha)p(\alpha)$.

B.2.1 Update of $q_Z^*(Z)$

$$\begin{aligned} \log q_Z^*(Z) &= \mathbb{E}_{\mathbf{W},\alpha}[\log p(I|Z) + \log p(Z|\mathbf{W})] + cst\,, \\ &\geqslant \mathbb{E}_{\mathbf{W},\alpha}[\log p(I|Z) + \log F(Z,\mathbf{W},\xi)]\,. \end{aligned} \tag{B.17}$$

Recalling that $p(I|Z) = \prod_n r_n^{Z_n}(1-r_n)^{1-Z_n}$, we have:

$$\mathbb{E}_{\mathbf{W},\alpha}[\log p(I|Z)] = \sum_n Z_n \log r_n + (1-Z_n)\log(1-r_n)\,. \tag{B.18}$$

The prior $p(Z|\mathbf{W}) = \prod_n [\sigma(w_n)]^{Z_n}[\sigma(-w_n)]^{1-Z_n}$ is lower bounded by $F(Z,\mathbf{W},\xi)$ to give:

$$\begin{aligned} \mathbb{E}_{\mathbf{W},\alpha}[\log F(Z,\mathbf{W},\xi)] = \sum_n Z_n[&\log\sigma(\xi_n) + (\mathbb{E}[w_n] - \xi_n)/2 \\ &- \lambda(\xi_n)(\mathbb{E}[w_n^2] - \xi_n^2)] + (1-Z_n)[\log\sigma(\xi_n) \\ &- (\mathbb{E}[w_n] + \xi_n)/2 - \lambda(\xi_n)(\mathbb{E}[w_n^2] - \xi_n^2)]\,. \end{aligned} \tag{B.19}$$

Summing (B.18) and (B.19) and taking the exponential, we have $q_Z^*(Z) \propto \prod_n \rho_{n0}^{1-Z_n}\rho_{n1}^{Z_n}$ where the expressions of ρ_{n0} and ρ_{n1} are given by (4.11). With the normalization constraint, we finally obtain $q_Z^*(Z) = \prod_n \eta_{n1}^{Z_n}\eta_{n0}^{1-Z_n}$ with $\eta_{nk} = \rho_{nk}/\sum_k \rho_{nk}$ for $k \in \{0,1\}$.

B.2.2 Update of $q_{\mathbf{W}}^*(\mathbf{W})$

$$\begin{aligned} \log q_{\mathbf{W}}^*(\mathbf{W}) &= \mathbb{E}_{Z,\alpha}[\log p(Z|\mathbf{W}) + \log p(\mathbf{W}|\alpha)] + cst\,, \\ &\geqslant \mathbb{E}_{Z,\alpha}[\log F(Z,\mathbf{W},\xi) + \log p(\mathbf{W}|\alpha)]\,. \end{aligned} \tag{B.20}$$

With the expression of $p(\mathbf{W}|\alpha)$ given in (4.5) and assuming that $q_{\mathbf{W}}(\mathbf{W}) = \prod_n q_{w_n}(w_n)$, we obtain:

$$\begin{aligned} \log\, q_{w_n}^*(w_n) = &-\frac{1}{2}\left[2\lambda(\xi_n)\left(w_n - \frac{1}{2\lambda(\xi_n)}\left(\eta_{n1} - \frac{1}{2}\right)\right)^2\right] \\ &-\frac{1}{2}\mathbb{E}_{w_j,\, j\neq n}\left[\frac{\alpha}{2}\sum_d (w_n - w_{\delta_d(n-2)})^2 + (w_n - w_{\delta_d(n+2)})^2\right]\,. \end{aligned} \tag{B.21}$$

By identifying the quadratic and linear terms in w_n, we obtain the formula for $\mathbf{\Sigma}_{w_n}$ and μ_{w_n} given in (4.12) and (4.13).

B.2.3 Update of $q^*_\alpha(\alpha)$

$$\begin{aligned}\log q^*_\alpha(\alpha) &= \mathbb{E}_{\mathbf{W}}[\log p(\mathbf{W}|\alpha)] + cst\,, \\ &= \mathbb{E}_{\mathbf{W}}\left[\frac{N}{2}\log\alpha - \frac{\alpha}{4}\sum_n\sum_{d=1}^{D}(w_{\delta_d(n+1)} - w_{\delta_d(n-1)})^2\right] + cst\,.\end{aligned} \tag{B.22}$$

Assuming $q^*_\alpha(\alpha)$ to be a Dirac distribution, we take the derivative of (B.22) with respect to α which leads to the update formula given in (4.14).

B.2.4 Update of $q^*_{\xi_n}(\xi_n)$

$$\begin{aligned}\log q^*_{\xi_n}(\xi_n) &= \mathbb{E}_{Z,\mathbf{W}}[\log F(Z,\mathbf{W},\xi)] + cst\,, \\ &= \log\sigma(\xi_n) - \frac{\xi_n}{2} - \lambda(\xi_n)(\mathbb{E}[w_n^2] - \xi_n^2) + cst\,.\end{aligned} \tag{B.23}$$

Taking the derivative with respect to ξ_n and setting it equal to zero gives $\lambda'(\xi_n)(\mathbb{E}[w_n^2] - \xi_n^2) = 0$. As $\lambda'(\xi_n) \geqslant 0$, we finally obtain the formula reported in (4.15).

B.3 FDSP prior - lower bound

The lower bound on the log-likelihood is used as a stopping criterion. To compute $\mathcal{J}(q)$, we need to evaluate the right hand side of (4.10):

$$\begin{aligned}\mathcal{J}(q) = \mathbb{E}[\log p(I|Z)] + \mathbb{E}[\log F(Z,\mathbf{W},\xi)] + \mathbb{E}[\log p(\mathbf{W}|\alpha)] \\ - \mathbb{E}[\log q_Z(Z)] - \mathbb{E}[\log q_{\mathbf{W}}(\mathbf{W})]\,.\end{aligned} \tag{B.24}$$

The values of the different expectations can be computed and are reported below.

$$\mathbb{E}[\log p(I|Z)] = \sum_n \eta_{n1}\log r_n + \eta_{n0}\log(1-r_n)\,. \tag{B.25}$$

$$\begin{aligned}\mathbb{E}[\log F(Z,\mathbf{W},\xi)] = \sum_n \eta_{n1}\mathbb{E}[w_n] + \log\sigma(\xi_n) \\ - \frac{\mathbb{E}[w_n] + \xi_n}{2} - \lambda(\xi_n)(\mathbb{E}[w_n^2] - \xi_n^2)\,.\end{aligned} \tag{B.26}$$

$$\mathbb{E}\left[\log p(\mathbf{W}|\alpha)\right] = \frac{N}{2}(\log\alpha - 1) + cst\,. \tag{B.27}$$

$$\mathbb{E}\left[\log q_Z(Z)\right] = \sum_n \eta_{n1} \log \eta_{n1} + \eta_{n0} \log \eta_{n0} \,. \tag{B.28}$$

$$\mathbb{E}\left[\log q_{\mathbf{W}}(\mathbf{W})\right] = -\frac{1}{2} \sum_n \log \mathbf{\Sigma}_{w_n} + cst \,. \tag{B.29}$$

Appendix C

Robust Bayesian fusion of continuous segmentation maps

Contents

C.1 Variational updates . . . 159
C.1.1 Robust probabilistic framework . . . 159
C.1.2 Mixture of consensuses . . . 163
C.2 Lower bound . . . 164
C.2.1 Robust probabilistic framework . . . 164
C.2.2 Mixture of consensuses . . . 166
C.3 Additional expectations . . . 166
C.3.1 Robust probabilistic model . . . 167
C.3.2 Mixture of consensuses . . . 167

C.1 Variational updates

Derivations of the variational update formula are given in this appendix, for the robust probabilistic framework of section 5.2 that uses heavy-tailed distributions, and for the mixture of consensuses model of section 5.3.

C.1.1 Robust probabilistic framework

The log joint probability of the heavy-tailed probabilistic model $p(\tilde{\mathbf{D}}, \tilde{\mathbf{T}}, \mathbf{b}, \mathbf{W}, S)$ factorizes as $p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, \mathbf{b}, S)p(\mathbf{b})p(\tilde{\mathbf{T}}|\mathbf{W})p(\mathbf{W})p(S)$.

Update of $q(\tilde{\mathbf{T}})$.

Eq. 5.16 applied to the consensus posterior approximation gives:

$$\begin{aligned} \log q^*(\tilde{\mathbf{T}}) &= \mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, \mathbf{b}, S) + \log p(\tilde{\mathbf{T}}|\mathbf{W})] + cst\,, \\ &= \mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, \mathbf{b}, \tau) + \log p(\tilde{\mathbf{T}}|\mathbf{W})] + cst\,. \end{aligned} \tag{C.1}$$

$p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, \mathbf{b}, \tau)$ is a Gaussian distribution according to the scale mixture representation. Discarding the terms independent of $\tilde{\mathbf{T}}_n$, $\mathbb{E}[\log p(\tilde{\mathbf{D}}_n|\tilde{\mathbf{T}}_n, \mathbf{b}, \tau_n)]$ is written:

$$\begin{aligned}\mathbb{E}[\log p(\tilde{\mathbf{D}}_n|\tilde{\mathbf{T}}_n, \mathbf{b}, \tau_n)] = -\frac{1}{2}\tilde{\mathbf{T}}_n^T \left(\sum_{p=1}^{P} \mathbf{\Sigma}_p^{-1}\mathbb{E}[\tau_n^p]\right) \tilde{\mathbf{T}}_n \\ + \tilde{\mathbf{T}}_n^T \left(\sum_{p=1}^{P} \mathbf{\Sigma}_p^{-1}\mathbb{E}[\tau_n^p](\tilde{\mathbf{D}}_n^p - \mathbb{E}[\mathbf{b}_p])\right) + cst\,.\end{aligned} \tag{C.2}$$

The second term in Eq. C.1 is due to the spatial regularization and can be expressed as follows:

$$\mathbb{E}[\log p(\tilde{\mathbf{T}}_n|\mathbf{W})] = -\frac{1}{2}\tilde{\mathbf{T}}_n^T \mathbf{\Sigma}_T^{-1}\mathbf{I}_K\tilde{\mathbf{T}}_n + \tilde{\mathbf{T}}_n^T\mathbf{\Sigma}_T^{-1}\mathbb{E}[\mathbf{W}]\mathbf{\Phi}_n + cst\,. \tag{C.3}$$

After regrouping and identifying the quadratic and linear terms in $\tilde{\mathbf{T}}_n$, we recognize a Gaussian distribution of parameters given by Eqs. 5.17 and 5.18.

Update of $q(\mathbf{b})$.

Following the same approach, we have for the rater bias:

$$\log q^*(\mathbf{b}) = \mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, \mathbf{b}, \tau) + \log p(\mathbf{b}|\beta)] + cst\,. \tag{C.4}$$

Considering rater p, the first term of Eq. C.4 gives:

$$\begin{aligned}\mathbb{E}[\log p(\tilde{\mathbf{D}}^p|\tilde{\mathbf{T}}, \mathbf{b}_p, \tau^p)] = -\frac{1}{2}\mathbf{b}_p^T \left(\sum_{n=1}^{N} \mathbf{\Sigma}_p^{-1}\mathbb{E}[\tau_n^p]\right) \mathbf{b}_p \\ + \mathbf{b}_p^T \left(\sum_{n=1}^{N} \mathbf{\Sigma}_p^{-1}\mathbb{E}[\tau_n^p](\tilde{\mathbf{D}}_n^p - \mathbb{E}[\tilde{\mathbf{T}}_n])\right) + cst\,,\end{aligned} \tag{C.5}$$

and the second term can be written as $\mathbb{E}[\log p(\mathbf{b}_p|\beta)] = -\frac{\beta}{2}\mathbf{b}_p^T\mathbf{b}_p + cst$. Combining the two and rearranging leads to the Gaussian distribution described by Eqs. 5.19 and 5.20.

Update of $q(\tau)$.

We now present the derivations for the posterior approximation of the scale factor τ. Discarding the terms independent of τ_n^p, Eq. 5.16 gives:

$$\log q^*(\tau_n^p) = \mathbb{E}[\log p(\tilde{\mathbf{D}}_n^p|\tilde{\mathbf{T}}, \mathbf{b}_p, \tau_n^p) + \log p(\tau_n^p)] + cst\,. \tag{C.6}$$

The results for the different distributions are reported below.

Student's *t*-distribution. The prior over the scale factor follows a Gamma distribution. Eq. C.6 can then be re-written as follows:

$$\begin{aligned}\log q^*(\tau_n^p) = &\left(\frac{K+\nu_p}{2} - 1\right)\log \tau_n^p \\ &- \frac{1}{2}\tau_n^p \left(\mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)^T \boldsymbol{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)] + \nu_p\right) + cst\,.\end{aligned} \tag{C.7}$$

We recognize a Gamma distribution of parameters $\frac{K+\nu_p}{2}$ and $\frac{\nu_p+E}{2}$ as given in Tab. 5.2, with $E = \mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)^T \boldsymbol{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)]$.

Laplace distribution. In this case, the prior is defined as an inverse Gamma distribution of parameters $\frac{K+1}{2}$ and $\frac{1}{8}$. Eq. C.6 leads to:

$$\begin{aligned}\log q^*(\tau_n^p) &= \frac{K}{2}\log \tau_n^p - \frac{\tau_n^p}{2}E - \frac{K+3}{2}\log \tau_n^p - \frac{1}{8\tau_n^p} + cst\,, \\ &= -\frac{3}{2}\log \tau_n^p - \frac{1}{2}\left(\frac{E}{\tau_n^p}\left(\tau_n^p - \frac{1}{\sqrt{4E}}\right)^2\right) + cst\,,\end{aligned} \tag{C.8}$$

where E is defined above.

Thus, the scale factor posterior approximation is an inverse Gaussian distribution whose parameters are given in Tab. 5.2.

GDP distribution. The prior over the scale factor is also an inverse Gamma distribution, but of parameters $\frac{K+1}{2}$ and $\frac{(z_n^p)^2}{2}$. Thus, we have:

$$\begin{aligned}\log q^*(\tau_n^p) &= \frac{K}{2}\log \tau_n^p - \frac{\tau_n^p E}{2} - \frac{K+3}{2}\log \tau_n^p - \frac{\mathbb{E}[(z_n^p)^2]}{2\tau_n^p} + cst\,, \\ &= -\frac{3}{2}\log \tau_n^p - \frac{1}{2}\left(\frac{E}{\tau_n^p}\left(\tau_n^p - \sqrt{\frac{\mathbb{E}[(z_n^p)^2]}{E}}\right)^2\right) + cst\,.\end{aligned} \tag{C.9}$$

Therefore, Eq. 5.16 again yields an inverse Gaussian distribution with the parameters given in Tab. 5.2.

Update of $q(z)$.

This section gives the derivations for the additional scale factor z which appears when the generalized double Pareto distribution is written as a Laplace scale mixture. Eq. 5.16 applied to $q(z_n^p)$ gives:

$$\begin{aligned}\log q^*(z_n^p) &= (K+1)\log z_n^p - \mathcal{T}_n^p \frac{(z_n^p)^2}{2} + (\nu_p - 1)\log z_n^p - \nu_p z_n^p + cst\,, \\ &= (K+\nu_p)\log z_n^p - \frac{1}{2}\left((z_n^p)^2 \mathcal{T}_n^p + 2\nu_p z_n^p\right) + cst\,,\end{aligned} \tag{C.10}$$

where $\mathcal{T}_n^p = \mathbb{E}\left[\frac{1}{\tau_n^p}\right]$. The normalization constant of Eq. 5.21 can be obtained by integration of Eq. C.10. Let $J^+(p,q,r)$ be the following integral:

$$J^+(p,q,r) = \int_0^\infty x^p \exp(qx - rx^2)dx\,, \tag{C.11}$$

with $p \geq 0$, $-\infty < q < \infty$ and $r > 0$.

It can then be shown [Neville, 2013] that:

$$J^+(p,q,r) = (2r)^{-\frac{p+1}{2}}\Gamma(p+1)\exp\left(\frac{q^2}{8r}\right)D_{-p-1}\left(-\frac{q}{\sqrt{2r}}\right)\,, \tag{C.12}$$

where D_ν is the parabolic cylinder function of order $\nu \in \mathbb{R}$.

From Eq. C.10, we have that $q^*(z_n^p) \propto (z_n^p)^{K+\nu_p}\exp\left(-\frac{(z_n^p)^2}{2}\mathcal{T}_n^p - \nu_p z_n^p\right)$. Therefore, using Eq. C.12 with $p = K+\nu_p$, $q = -\nu_p$ and $r = \frac{\mathcal{T}_n^p}{2}$, we get the density of Eq. 5.21. In practice, we only need $\mathbb{E}[z_n^p]$ and $\mathbb{E}[(z_n^p)^2]$ to perform the inference. These expectations can be computed using the same approach. For $\mathbb{E}[z_n^p]$, we use Eq. C.12 with $p = K+\nu_p+1$, $q = -\nu_p$, and $r = \frac{\mathcal{T}_n^p}{2}$ which gives:

$$\mathbb{E}[z_n^p] = \frac{(K+\nu_p+1)D_{-K-\nu_p-2}\left(\frac{\nu_p}{\sqrt{\mathcal{T}_n^p}}\right)}{\sqrt{\mathcal{T}_n^p}D_{-K-\nu_p-1}\left(\frac{\nu_p}{\sqrt{\mathcal{T}_n^p}}\right)}\,. \tag{C.13}$$

Likewise, we have $p = K+\nu_p+2$ with same q and r for $\mathbb{E}[(z_n^p)^2]$, which yields:

$$\mathbb{E}[(z_n^p)^2] = \frac{(K+\nu_p+1)(K+\nu_p+2)D_{-K-\nu_p-3}\left(\frac{\nu_p}{\sqrt{\mathcal{T}_n^p}}\right)}{\mathcal{T}_n^p D_{-K-\nu_p-1}\left(\frac{\nu_p}{\sqrt{\mathcal{T}_n^p}}\right)}\,. \tag{C.14}$$

The function $R_\nu(x)$ defined as $R_\nu(x) = \frac{D_{-\nu-2}(x)}{D_{-\nu-1}}(x)$ leads to underflow problems for large x or ν. Therefore, we follow [Neville, 2013] and compute the ratio using Lentz's algorithm, which is based on the continued fraction representation of the function.

Update of $q(\mathbf{W})$.

Eq. 5.16 applied to $\mathbf{W}_k$ gives:

$$\begin{aligned}\log q^*(\mathbf{W}_k) &= \mathbb{E}[\log p(\tilde{\mathbf{T}}_k|\mathbf{W}_k) + \log p(\mathbf{W}_k)] + cst\,,\\ &= -\frac{1}{2}\mathbf{W}_k^T\left(\mathbf{\Sigma}_T^{-1}\sum_{n=1}^N \mathbf{\Phi}_n^T\mathbf{\Phi}_n\right)\mathbf{W}_k\\ &\quad + \mathbf{W}_k^T\left(\mathbf{\Sigma}_T^{-1}\sum_{n=1}^N \mathbf{\Phi}_n\mathbb{E}[\tilde{\mathbf{T}}_{nk}]\right) - \frac{\alpha}{2}\mathbf{W}_k^T\mathbf{W}_k + cst\,.\end{aligned} \tag{C.15}$$

Regrouping the quadratic and linear terms in $\mathbf{W}_k$, we obtain a Gaussian distribution whose parameters are given by Eqs. 5.22 and 5.23.

Update of the remaining parameters.

The update formulas for α, $\mathbf{\Sigma}_T$, $\mathbf{\Sigma}_p$, β and ν_p are obtained by considering these parameters as variables and assuming that their posterior approximation $q(.)$ is a Dirac distribution. The mode of the posterior distribution approximation is found by maximizing Eq. 5.16, which leads to Eqs. 5.24–5.28.

C.1.2 Mixture of consensuses

The model log joint probability $p(\tilde{\mathbf{D}}, \tilde{\mathbf{T}}, Z)$ factorizes as $p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, Z)p(Z)$.

Update of $q(Z)$.

Eq. 5.16 applied to the variable Z leads to:

$$\log q^*(Z) = \mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, Z) + \log p(Z)] + cst\,. \tag{C.16}$$

The first term can be developed to give:

$$\begin{aligned}\mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, Z)] = \sum_{p=1}^{P}\sum_{m=1}^{M} z_{pm}\Bigg(\sum_{n=1}^{N} -\frac{K}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_p| \\ - \frac{1}{2}\mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})^T\mathbf{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})]\Bigg)\,,\end{aligned} \tag{C.17}$$

and the second term is equal to $\sum_{p=1}^{P}\sum_{m=1}^{M} z_{pm}\log\pi_m$.

Summing the two expectations and taking the exponential, we get $q^*(Z) \propto \prod_p\prod_m \rho_{pm}^{Z_{pm}}$, where the expression of ρ_{pm} is given by Eq. 5.30. We finally obtain a product of categorical distributions with parameters r_{pm}, such that $q^*(Z) = \prod_p\prod_m r_{pm}^{Z_{pm}}$.

Update of $q(\tilde{\mathbf{T}})$.

Applying Eq. 5.16 to the consensus posterior approximation and discarding the terms independent with respect to the mth map leads to:

$$\log q^*(\tilde{\mathbf{T}}_{nm}) = -\frac{1}{2}\tilde{\mathbf{T}}_{nm}^T\left(\sum_{p=1}^{P} r_{pm}\mathbf{\Sigma}_p^{-1}\right)\tilde{\mathbf{T}}_{nm} + \tilde{\mathbf{T}}_{nm}^T\left(\sum_{p=1}^{P} r_{pm}\mathbf{\Sigma}_p^{-1}\tilde{\mathbf{D}}_n^p\right) + cst\,. \tag{C.18}$$

We recognize a Gaussian distribution whose parameters are given by Eq. 5.31 and Eq. 5.32.

Update of the model parameters.

The update formulas for $\mathbf{\Sigma}_p$ and π_m are obtained by considering them as variables whose approximate posterior is a Dirac distribution. We find the mode of each distribution by maximizing Eq. 5.16 and using the fact that $\sum_m \pi_m = 1$ for the mixing coefficients, which leads to Eqs. 5.33 and 5.34.

C.2 Lower bound

In this chapter, we propose a variational inference scheme to estimate the posterior approximations. It is based on the maximization of a lower bound $\mathcal{L}(q)$ over the data marginal log likelihood. The lower bound can be computed and is used in practice as a stopping criterion, except when the framework is based on a GDP distribution, because of the long computation time.

C.2.1 Robust probabilistic framework

We can re-write Eq. 5.15 as follows:

$$\begin{aligned}\mathcal{L}(q) = &\mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, \mathbf{b}, S)] + \mathbb{E}[\log p(S)] + \mathbb{E}[\log p(\mathbf{b})] \\ &+ \mathbb{E}[\log p(\tilde{\mathbf{T}}|\mathbf{W})] + \mathbb{E}[\log p(\mathbf{W})] \\ &- \mathbb{E}[\log q(S)] - \mathbb{E}[\log q(\mathbf{b})] - \mathbb{E}[\log q(\tilde{\mathbf{T}})] - \mathbb{E}[\log q(\mathbf{W})]\,.\end{aligned} \tag{C.19}$$

The values of the different expectations are reported bellow.

Expectations involving the scale factors.

We first focus on the expectations involving the scale factors τ and z. $\mathbb{E}[\log p(\tau_n^p)]$ is given in Tab. C.1 and $\mathbb{E}[\log q(\tau_n^p)]$ is given in Tab. C.2.

Tab. C.1: Formula giving $\mathbb{E}[\log p(\tau_n^p)]$ for the three heavy-tailed likelihoods. The values of the constants for the Laplace and GDP distributions are given by $C_{\mathrm{L}} = -\frac{K+1}{2}\log 8 - \log\Gamma\left(\frac{K+1}{2}\right)$ and $C_{\mathrm{GDP}} = -\frac{K+1}{2}\log 2 - \log\Gamma\left(\frac{K+1}{2}\right)$, respectively.

Likelihood	$\mathbb{E}[\log p(\tau_n^p)]$	
Student's t	$-\log\Gamma\left(\frac{\nu_p}{2}\right) + \frac{\nu_p}{2}\log\frac{\nu_p}{2} + \left(\frac{\nu_p}{2}-1\right)\mathbb{E}[\log\tau_n^p] - \frac{\nu_p}{2}\mathbb{E}[\tau_n^p]$	(C.20)
Laplace	$-\frac{K+3}{2}\mathbb{E}[\log\tau_n^p] - \frac{1}{8}\mathbb{E}\left[\frac{1}{\tau_n^p}\right] + C_{\mathrm{L}}$	(C.21)
GDP	$(K+1)\mathbb{E}[\log z_n^p] - \frac{K+3}{2}\mathbb{E}[\log\tau_n^p] - \frac{1}{2}\mathbb{E}[(z_n^p)^2]\mathbb{E}\left[\frac{1}{\tau_n^p}\right] + C_{\mathrm{GDP}}$	(C.22)

Tab. C.2: Formula giving $\mathbb{E}[\log q(\tau_n^p)]$ for the three heavy-tailed likelihoods. a_{np} and b_{np} are given in Tab. 5.2.

Likelihood	$\mathbb{E}[\log q(\tau_n^p)]$	
Student's t	$-\log\Gamma(a_{np}) + a_{np}\log b_{np} + (a_{np}-1)\mathbb{E}[\log\tau_n^p] - b_{np}\mathbb{E}[\tau_n^p]$	(C.23)
Laplace	$-\frac{3}{2}\mathbb{E}[\log\tau_n^p] - \frac{1}{2}\log(8\pi) - \frac{1}{2}$	(C.24)
GDP	$\frac{1}{2}\log\frac{\mathbb{E}[(z_n^p)^2]}{2\pi} - \frac{3}{2}\mathbb{E}[\log\tau_n^p] - \frac{1}{2}$	(C.25)

For the GDP likelihood, there is the additional latent variable z. Expectations involving this scale factor are given below.

$$\mathbb{E}[\log p(z_n^p|\nu_p)] = (\nu_p - 1)\mathbb{E}[\log z_n^p] + \nu_p \log \nu_p - \log \Gamma(\nu_p) - \nu_p \mathbb{E}[z_n^p]\,. \tag{C.26}$$

$$\begin{aligned}\mathbb{E}[\log q(z_n^p)] = \frac{K+\nu_p+1}{2}\log \mathcal{T}_n^p + (K+\nu_p)\mathbb{E}[\log z_n^p] - \nu_p \mathbb{E}[z_n^p] \\ - \frac{\mathcal{T}_n^p}{2}\mathbb{E}[(z_n^p)^2] - \log\Gamma(K+\nu_p+1) - \frac{\nu_p^2}{4\mathcal{T}_n^p} - \log D_{-K-\nu_p-1}\left(\frac{\nu_p}{\sqrt{\mathcal{T}_n^p}}\right)\,,\end{aligned} \tag{C.27}$$

where $\mathcal{T}_n^p = \mathbb{E}\left[\frac{1}{\tau_n^p}\right]$.

Appendix C.1 explains how to compute $\mathbb{E}[z_n^p]$ and $\mathbb{E}[(z_n^p)^2]$. The expectation $\mathbb{E}[\log z_n^p]$ vanishes when computing $\mathbb{E}[\log p(\tau_n^p)] + \mathbb{E}[\log p(z_n^p)] - \mathbb{E}[\log q(z_n^p)]$ with Eq. C.22, C.26 and C.27 and it does not need to be evaluated in practice. Other expectations are given in appendix C.3.

Remaining expectations.

$$\begin{aligned}\mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}},\tau,\mathbf{b})] = \sum_{n=1}^{N}\sum_{p=1}^{P}\Bigg(-\frac{1}{2}\log|\mathbf{\Sigma}_p| + \frac{K}{2}\mathbb{E}[\log\tau_n^p] \\ - \frac{1}{2}\Big[(\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p})^T\mathbb{E}[\tau_n^p]\mathbf{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p}) \\ + \mathbb{E}[\tau_n^p]\Big(\operatorname{Tr}(\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_{\mathbf{b}_p}) + \operatorname{Tr}(\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n})\Big)\Big]\Bigg) + cst\,.\end{aligned} \tag{C.28}$$

$$\begin{aligned}\mathbb{E}[\log p(\tilde{\mathbf{T}}|\mathbf{W})] = \sum_{n=1}^{N}\Bigg(-\frac{K}{2}\log\mathbf{\Sigma}_T - \frac{1}{2}\Big[(\mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{W}}\mathbf{\Phi}_n)^T\mathbf{\Sigma}_T^{-1}\mathbf{I}_K(\mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{W}}\mathbf{\Phi}_n) \\ + \mathbf{\Sigma}_T^{-1}\operatorname{Tr}(\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n}) + \mathbf{\Sigma}_T^{-1}\sum_{k=1}^{K}\operatorname{Tr}(\mathbf{\Phi}_n\mathbf{\Phi}_n^T\mathbf{\Sigma}_{\mathbf{W}_k})\Big]\Bigg) + cst\,.\end{aligned} \tag{C.29}$$

$$\mathbb{E}[\log p(\mathbf{b})] = \sum_{p=1}^{P}\frac{K}{2}\log\beta - \frac{\beta}{2}\left[\mu_{\mathbf{b}_p}^T\mu_{\mathbf{b}_p} + \operatorname{Tr}(\mathbf{\Sigma}_{\mathbf{b}_p})\right] + cst\,. \tag{C.30}$$

$$\mathbb{E}[\log p(\mathbf{W})] = \sum_{k=1}^{K}\frac{L}{2}\log\alpha - \frac{\alpha}{2}\left[\mu_{\mathbf{W}_k}^T\mu_{\mathbf{W}_k} + \operatorname{Tr}(\mathbf{\Sigma}_{\mathbf{W}_k})\right] + cst\,. \tag{C.31}$$

Finally, $\mathbb{E}[\log q(\tilde{\mathbf{T}})]$, $\mathbb{E}[\log q(\mathbf{b})]$ and $\mathbb{E}[\log q(\mathbf{W})]$ are given by:

$$\mathbb{E}[\log q(\tilde{\mathbf{T}})] = \sum_{n=1}^{N} -\frac{1}{2}\log|\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n}| + cst\,, \tag{C.32}$$

$$\mathbb{E}[\log q(\mathbf{b}_p)] = \sum_{p=1}^{P} -\frac{1}{2}\log|\mathbf{\Sigma}_{\mathbf{b}_p}| + cst\,, \tag{C.33}$$

$$\mathbb{E}[\log q(\mathbf{W})] = \sum_{k=1}^{K} -\frac{1}{2}\log|\mathbf{\Sigma}_{\mathbf{W}_k}| + cst\,. \tag{C.34}$$

C.2.2 Mixture of consensuses

The lower bound for the mixture of consensuses model can be written as follows:

$$\mathcal{L}(q) = \mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, Z)] + \mathbb{E}[\log p(Z)] - \mathbb{E}[\log q(\tilde{\mathbf{T}})] - \mathbb{E}[\log q(Z)]\,. \tag{C.35}$$

Developing each term, we obtain:

$$\begin{aligned}\mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, Z)] = \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{p=1}^{P} r_{pm}\Big[&-\frac{K}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}_p| \\ &- \frac{1}{2}\Big((\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{nm}})^T\mathbf{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{nm}}) + \mathrm{Tr}(\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}})\Big)\Big]\,,\end{aligned} \tag{C.36}$$

$$\mathbb{E}[\log p(Z)] = \sum_{p=1}^{P}\sum_{m=1}^{M} r_{pm}\log\pi_m\,, \tag{C.37}$$

$$\mathbb{E}[\log q(\tilde{\mathbf{T}})] = \sum_{n=1}^{N}\sum_{m=1}^{M} -\frac{1}{2}\log|\mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}}| + cst\,, \tag{C.38}$$

$$\mathbb{E}[\log q(Z)] = \sum_{p=1}^{P}\sum_{m=1}^{M} r_{pm}\log r_{pm}\,. \tag{C.39}$$

C.3 Additional expectations

In this last section, we gather together some useful expectations involved in the variational updates or in the evaluation of the lower bound.

C.3.1 Robust probabilistic model

$$\mathbb{E}[\mathbf{b}_p] = \mu_{\mathbf{b}_p} \,. \tag{C.40}$$

$$\mathbb{E}[\tilde{\mathbf{T}}_n] = \mu_{\tilde{\mathbf{T}}_n} \,. \tag{C.41}$$

$$\mathbb{E}[\tilde{\mathbf{T}}_{nk}] = \mu_{\tilde{\mathbf{T}}_{nk}} \,. \tag{C.42}$$

$$\mathbb{E}[\mathbf{W}_k] = \mu_{\mathbf{W}_k} \,. \tag{C.43}$$

$$\begin{aligned}\mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)^T \boldsymbol{\Sigma}_p^{-1} (\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)] = \\ (\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p})^T \boldsymbol{\Sigma}_p^{-1} (\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p}) + \mathrm{Tr}(\boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}_{\mathbf{b}_p}) + \mathrm{Tr}(\boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}_{\tilde{\mathbf{T}}_n}) \,.\end{aligned} \tag{C.44}$$

Moreover, $\mathbb{E}[\mathbf{W}]$ corresponds to the gathering of the expectations $\mathbb{E}[\mathbf{W}_k]$ given above in a matrix of size $K \times L$.

Regarding the scale factor τ, we need the expectations $\mathbb{E}[\tau_n^p]$ and $\mathbb{E}[\log \tau_n^p]$ for the Student's t-distribution. There are given by:

$$\mathbb{E}[\tau_n^p] = \frac{a_{np}}{b_{np}} \,, \tag{C.45}$$

$$\mathbb{E}[\log \tau_n^p] = \psi(a_{np}) - b_{np} \,, \tag{C.46}$$

where a_{np} and b_{np} are the parameters of the Gamma distribution described in Tab. 5.2 and ψ is the digamma function. For the Laplace and GDP distributions, we have to evaluate $\mathbb{E}[\tau_n^p]$ and $\mathbb{E}\left[\frac{1}{\tau_n^p}\right]$. The latter is notably involved in the estimation of the degree of freedom. They can be written as follows:

$$\mathbb{E}[\tau_n^p] = \mu_{np} \,, \tag{C.47}$$

$$\mathbb{E}\left[\frac{1}{\tau_n^p}\right] = \frac{1}{\mu_{np}} + \frac{1}{\lambda_{np}} \,, \tag{C.48}$$

where μ_{np} and λ_{np} are the parameters of the inverse Gaussian distributions given in Tab. 5.2.

Moreover, a third expectation, $\mathbb{E}[\log \tau_n^p]$, appears in some terms of the lower bound. In contrast to the Gaussian case, it does not have a closed-form formula for the Laplace or GDP distributions. However, this is not a problem in practice as it vanishes when gathering the different parts, in particular after summation of Eqs. C.21, C.24 and C.28 for the Laplace distribution, and summation of Eqs. C.22, C.25 and C.28 for the GDP distribution.

C.3.2 Mixture of consensuses

$$\mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})^T \boldsymbol{\Sigma}_p^{-1} (\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})] = (\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{nm}})^T \boldsymbol{\Sigma}_p^{-1} (\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{nm}}) + \mathrm{Tr}(\boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}_{\tilde{\mathbf{T}}_{nm}}) \,. \tag{C.49}$$

Appendix D

End-to-end analysis of a computerized lung cancer screening pipeline based on LDCT

Contents

D.1 Circulant miRNAs as potential biomarkers for lung cancer detection . . . 169

D.1.1 Introduction . . . 169

D.1.2 Data collection and pre-processing . . . 170

D.1.3 Experiments on the COSMOS dataset . . . 171

D.1.4 Conclusion . . . 173

D.1 Circulant miRNAs as potential biomarkers for lung cancer detection

D.1.1 Introduction

A key factor of lung cancer prognosis is the time of detection, as chances of survival drop drastically for late stage diagnostic. The implementation of large scale screening programs requires the development of markers for lung cancer early detection. In this chapter, we focused on imaging markers, whose effectiveness has been demonstrated by large scale screening trials, leading LDCT to become the cornerstone of the United States lung cancer screening policy.

However, LDCT screening has also some limitations, including the substantial cost of the equipment required for the image acquisition, the challenges related to the analysis of the scans, the high rate of false positives, but also the concerns about the impact of long term radiation exposure [Wood et al., 2018; McCunney & Li, 2014]. In contrast, blood-derived biomarkers, such as circulant miRNAs, can be easily collected by non-invasive liquid biopsy.

miRNAs are very short (about 20 nucleotides) non coding and single-stranded RNA molecules. Their synthesis begins in the nucleii of the cells and requires the intervention

of an RNA polymerase. The miRNAs precursors are subsequently transported to the cytoplasm where they acquire their mature form [Iqbal et al., 2019].

They play an important role in gene expression regulation by binding to the target region of some messenger RNAs (mRNAs), which induces their degradation or inhibits the translation, leading to post-transcriptional gene silencing. However, miRNAs are also suspected to contribute to the activation of the expression of some genes, either directly or indirectly.

In recent years, the profiling of miRNA expression has been an active topic of research, as many miRNAs have been reported as being associated with human diseases, suggesting their potential as new biomarkers. In particular, some miRNAs exhibit enhanced expression levels in lung cancer cases. Some appear to be specifically related to certain subtypes of lung cancer, or to be involved in the differentiation process between subtypes. Others seem to contribute to the regulation of tumor suppressor genes, found to be inactivated in some lung cancer cases [Iqbal et al., 2019].

Various factors, such as mutations, epigenetic modifications, transcriptional repression or defective biogenesis, are suggested to explain the changes in expression level observed for some miRNAs in cancer cases. Interestingly, this deregulation can also be detected in circulating miRNAs. These molecules, also denoted as cell-free miRNAs, are circulating in the body fluids in exosomes, which are small membrane vesicles involved in the cell-to-cell communication [Fortunato et al., 2019]. Exosomes can be realeased by different cell types, including within the tumor environment. The attractive property of circulating miRNAs is their simple collection by non-invasive liquid biopsy which make them particularly suitable for the screening of diseases and the monitoring of their progression. This potential has been investigated by a large number of studies for different types of cancer, including breast cancer [Madhavan et al., 2016], gastric cancer [Huang et al., 2017], colorectal cancer [Huang et al., 2010] and lung cancer [Montani et al., 2015; Shen et al., 2011; Sozzi et al., 2014].

D.1.2 Data collection and pre-processing

A common approach to assess the concentration of a specific miRNA in a blood sample is the reverse transcriptase quantitative polymerase chain reaction (RT-qPCR). The serum, obtained after clotting and centrifugation, is purified to isolate and concentrate the ARN molecules. The latter are then converted in complementary DNA (cDNA) by a reverse transcriptase.

A PCR reaction is used to amplify the DNA sequences with target-specific primers labelled with markers emitting fluorescence only after hybridization. Real-time follow-up of the fluorescence intensity enables to estimate the sample DNA quantity, which doubles theoretically at each PCR cycle during the exponential phase. However, in practice, this depends on the PCR reaction efficiency, denoted as E, leading to the following equation:

$$X_n = X_0(1+E)^n\,, \tag{D.1}$$

where X_n is the DNA quantity at cycle n. The cycle threshold, denoted as C_T, is the number of PCR cycles required by a given miRNA to reach a pre-defined fluorescence intensity value. It is inversely proportional to the input copy number, as small initial

concentrations lead to larger C_T values, and conversely. Exemples of RT-qPCR curves are shown in Fig. D.1.

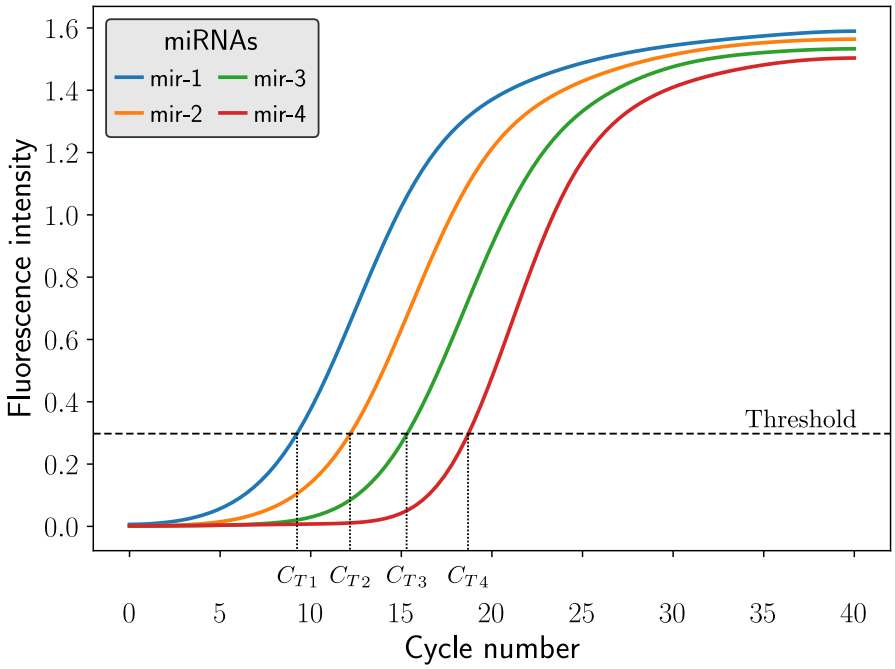


Fig. D.1: Fluorescence intensity curves for 4 miRNAs by RT-qPCR. ł

In practice, two methods are used to analyse RT-qPCR results. Absolute quantification allows the initial input copy number to be estimated, but requires the obtention of standard curves by dilution. In contrast, relative quantification compares the C_T values between samples. However, normalization is necessary to remove biases due to the quality of samples and the experimental conditions.

Many data normalization strategies have been proposed [Meng et al., 2017; Cheng et al., 2016; Schwarzenbach et al., 2015], but there is still no consensus about the optimal method. A common approach uses a reference gene r to assess the concentration of a target t. Reference genes, also denoted as housekeeping genes, are genes selected for their expression stability across the entire dataset and with respect to the experimental parameters. When the fluorescence curve reaches the intensity threshold, the following equality holds between the target and the reference: $X_0^t(1+E)^{C_{Tt}} = X_0^r(1+E)^{C_{Tr}}$, where we assume identical PCR reaction efficiencies. This leads to:

$$\frac{X_0^t}{X_0^r} = (1+E)^{-\Delta C_T}\,, \tag{D.2}$$

where $\Delta C_T = C_{Tt} - C_{Tr}$. Biomarker candidates are miRNAs presenting significant differences in ΔC_T values between cancer cases and controls.

D.1.3 Experiments on the COSMOS dataset

We now present some results obtained on the publicly available data of the Continuous Observation of Smoking Subjects (COSMOS) study, a lung cancer screening trial

Tab. D.1: 5-fold cross-validation results for different classifiers on the COSMOS data.

	Precision (%)	Recall (%)	F1-score (%)	AUC
LDA	**88.8** (±6.4)	73.3 (±13.1)	**79.0** (±6.6)	0.97 (±0.02)
QDA	85.3 (±8.4)	66.6 (±7.3)	74.5 (±5.2)	0.95 (±0.04)
Logistic regression	83.3 (±9.6)	67.3 (±18.3)	72.8 (±10.3)	0.98 (±0.02)
SVM	78.6 (±9.8)	**79.1** (±9.5)	77.8 (±4.2)	**0.98** (±0.01)
RVM_f [1]	76.4 (±7.3)	62.9 (±12.8)	67.9 (±8.5)	**0.98** (±0.01)
RVM_s [2]	84.6 (±12.2)	64.4 (±8.5)	72.7 (±8.2)	**0.98** (±0.02)

[1] Relevance Vector Machine - feature selection.
[2] Relevance Vector Machine - sample selection.

conducted in Italy and involving high-risk subjects above 50 years old with a smoking history [Montani et al., 2015].

This cohort was augmented with lung cancer patients diagnosed outside of the study, and then divided into a calibration and a validation sets of 24 and 1008 subjects, respectively. The total number of lung cancer cases is 48. The expressions of 34 miRNAs were quantified by RT-qPCR on the calibration set. Normalization with respect to 6 housekeeping genes leads to the definition of a signature of 13 miRNAs whose expression levels were found significantly different between the control and cancer cases. This signature was used to build a Diagonal Linear Discriminant Analysis (DLDA) classifier for lung cancer prediction on the calibration set. This model was then independently evaluated on the validation set, leading to an AUC, precision, recall and F1 scores of 0.85, 0.38, 0.78 and 0.51, respectively.

The normalized C_T values of the 13-miRNA signature were made publicly available for both sets and were used in this section to compare different classifiers, including the Linear Discriminant Analysis (LDA), the Quadratic Discriminant Analysis (QDA), the logistic regression with a L_1 penalty, the Support Vector Machine (SVM) and the Relevance Vector Machine (RVM) introduced in [Tipping & Faul, 2003]. Results obtained by 5-fold stratified cross-validation are presented in Tab. D.1.

The regularization parameters of the logistic regression and SVM were optimized by grid-search on each training set. The SVM is fitted with a squared exponential kernel. The logistic regression and the RVM are sparse methods allowing the most relevant features to be identified, leading to an additional selection among the 13 miRNAs. Moreover, the RVM fitted with a squared exponential kernel can also be used to perform sample selection, i.e. to select the most relevant subjects for the classification task. Denoted as relevant vectors, they are the counterpart of the SVM support vectors.

First, one can observe that all methods are equivalent with respect to the AUC criterion. However, there are discrepancies for the F1-score. The reason is that the dataset is heavily imbalanced with 1008 controls and only 48 cancer cases, a situation where the AUC score tends to be over-optimistic [Davis & Goadrich, 2006]. In contrast, the F1-score, which is the harmonic mean of precision and recall, enables to focus on the performances of the classifiers on the minority class [He & Garcia, 2009].

The best F1-scores are obtained with the LDA and SVM classifiers. LDA is the simplest of the investigated models, and assumes Gaussian distributions for both classes with a shared covariance matrix. In contrast, there is no such assumption of identical covariance matrix in QDA analysis, but the model leads to poorer performances.

SVM gives close results to LDA, with 46 subjects out of 856 selected on average on each training set. The RVM with a squared exponential kernel leads to an even sparser solution, with 6 relevant vectors in average. These two methods allow typical subjects for both classes to be identified.

Finally, the logistic regression and the RVM for feature selection give the poorest results, but they perform a further selection among the 13-miRNA signature. On average, the logistic regression and RVM decision rules use 11 and 6 miRNAs, respectively. Being able to identify the most relevant miRNAs is an important problem in the perspective of real life applications, as screening is easier to implement with a reduced number of molecules to be tested. It would have been interesting to evaluate the RVM and the logistic regression on the initial set of 34 miRNAs, in order to compare their miRNA selections to the current signature, but this set was not made publicly available.

D.1.4 Conclusion

Circulant miRNAs are promising biomarkers for lung cancer detection because of their easy collection and good prediction results. However, several limitations need to be addressed. First, there is a lack of reproducibility between studies, because of the absence of universally accepted protocol for the miRNA extraction, and because of the problems related to the data pre-processing, including the choice of normalization strategy. External validations on larger datasets are also required before real life applications.

Bibliography

[Adalsteinsson & Sethian, 1995] D. Adalsteinsson and J. A. Sethian. "A Fast Level Set Method for Propagating Interfaces". In: *Journal of Computational Physics* 118.2 (1995), pp. 269–277 (cit. on p. 33).

[Akhondi-Asl et al., 2014] A. Akhondi-Asl, L. Hoyte, M. E. Lockhart, and S. K. Warfield. "A Logarithmic Opinion Pool Based STAPLE Algorithm for the Fusion of Segmentations With Associated Reliability Weights". In: *IEEE Transactions on Medical Imaging* 33.10 (2014), pp. 1997–2009 (cit. on p. 75).

[Akhondi-Asl & Warfield, 2013] A. Akhondi-Asl and S. K. Warfield. "Simultaneous Truth and Performance Level Estimation Through Fusion of Probabilistic Segmentations". In: *IEEE Transactions on Medical Imaging* 32.10 (2013), pp. 1840–1852 (cit. on pp. 75, 78, 93).

[Ambroise et al., 1997] C. Ambroise, M. Dang, and G. Govaert. "Clustering of Spatial Data by the EM Algorithm". In: *geoENV I — Geostatistics for Environmental Applications*. Ed. by A. Soares, J. Gómez-Hernandez, and R. Froidevaux. Dordrecht: Springer Netherlands, 1997, pp. 493–504 (cit. on p. 32).

[Ambroise & Govaert, 1998] C. Ambroise and G. Govaert. "Convergence of an EM-type algorithm for spatial clustering". In: *Pattern Recognition Letters* 19.10 (1998), pp. 919–927 (cit. on p. 55).

[Arbelle et al., 2019] A. Arbelle, E. Elul, and T. R. Raviv. *QANet – Quality Assurance Network for Image Segmentation*. 2019. arXiv: 1904.08503 [cs.CV] (cit. on p. 47).

[Archambeau & Verleysen, 2007] C. Archambeau and M. Verleysen. "Robust Bayesian clustering". In: *Neural Networks* 20.1 (2007), pp. 129–138 (cit. on pp. 51, 76).

[Ardila et al., 2019] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, et al. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography". In: *Nature Medicine* 25.6 (June 2019), pp. 954–961 (cit. on pp. 108–109, 129).

[Armato III et al., 2004] S. G. Armato III, G. McLennan, M. F. McNitt-Gray, C. R. Meyer, D. Yankelevitz, et al. "Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community". In: *Radiology* 232.3 (2004), pp. 739–748 (cit. on p. 110).

[Armato III et al., 2011] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, et al. "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans". In: *Medical Physics* 38.2 (2011), pp. 915–931 (cit. on pp. 4, 34, 57, 88, 106–107).

[Arslan, 2004] O. Arslan. "Family of multivariate generalized t distributions". In: *Journal of Multivariate Analysis* 89.2 (2004), pp. 329–337 (cit. on p. 79).

[Asman & Landman, 2011] A. J. Asman and B. A. Landman. "Robust Statistical Label Fusion Through Consensus Level, Labeler Accuracy, and Truth Estimation (COLLATE)". In: *IEEE Transactions on Medical Imaging* 30.10 (2011), pp. 1779–1794 (cit. on p. 75).

[Asman & Landman, 2012] A. J. Asman and B. A. Landman. "Formulating Spatially Varying Performance in the Statistical Fusion Framework". In: *IEEE Transactions on Medical Imaging* 31.6 (June 2012), pp. 1326–1336 (cit. on p. 75).

[Asman & Landman, 2013] A. J. Asman and B. A. Landman. "Non-local statistical label fusion for multi-atlas segmentation". In: *Medical Image Analysis* 17.2 (2013), pp. 194–208 (cit. on p. 75).

[Audelan et al., 2020] B. Audelan, D. Hamzaoui, S. Montagne, R. Renard-Penna, and H. Delingette. "Robust Fusion of Probability Maps". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Cham: Springer International Publishing, 2020, pp. 259–268 (cit. on pp. 6, 74, 76).

[Audelan et al., 2021] B. Audelan, S. Lopez, P. Fillard, Y. Diascorn, B. Padovani, et al. "Validation of lung nodule detection a year before diagnosis in NLST dataset based on a deep learning system". In: *Submitted to ERS International Congress 2021* (2021) (cit. on pp. 6, 106).

[Audelan & Delingette, 2019] B. Audelan and H. Delingette. "Unsupervised Quality Control of Image Segmentation Based on Bayesian Learning". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, et al. Cham: Springer International Publishing, 2019, pp. 21–29 (cit. on pp. 5, 17, 23, 46, 49–50, 52, 55).

[Audelan & Delingette, 2020] B. Audelan and H. Delingette. "Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model". In: *Medical Image Analysis* (2020) (cit. on pp. 17, 22).

[Audelan & Delingette, 2021] B. Audelan and H. Delingette. "Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model". In: *Medical Image Analysis* 68 (2021), p. 101895 (cit. on pp. 5, 46).

[Babacan et al., 2008] S. D. Babacan, R. Molina, and A. K. Katsaggelos. "Parameter Estimation in TV Image Restoration Using Variational Distribution Approximation". In: *IEEE Transactions on Image Processing* 17.3 (2008), pp. 326–339 (cit. on pp. 17, 21, 28, 101).

[Babacan et al., 2009] S. D. Babacan, R. Molina, and A. K. Katsaggelos. "Variational Bayesian Blind Deconvolution Using a Total Variation Prior". In: *IEEE Transactions on Image Processing* 18.1 (2009), pp. 12–26 (cit. on pp. 17, 21–22, 28).

[Bai et al., 2015] J. Bai, X. Huang, S. Liu, Q. Song, and R. Bhagalia. "Learning orientation invariant contextual features for nodule detection in lung CT scans". In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. 2015, pp. 1135–1138 (cit. on p. 107).

[Barta et al., 2019] J. Barta, C. Powell, and J. Wisnivesky. "Global Epidemiology of Lung Cancer". In: *Annals of Global Health* 85.1 (2019), p. 8 (cit. on p. 106).

[Benzaquen et al., 2019] J. Benzaquen, J. Boutros, C. Marquette, H. Delingette, and P. Hofman. "Lung Cancer Screening, towards a Multidimensional Approach: Why and How?" In: *Cancers* 11.2 (2019) (cit. on p. 141).

[Betts et al., 2013] J. G. Betts, K. A. Young, J. A. Wise, E. Johnson, B. Poe, et al. *Anatomy and Physiology*. Houston, Texas: OpenStax, 2013 (cit. on p. 10).

[Bioucas-Dias & Figueiredo, 2016] J. M. Bioucas-Dias and M. A. T. Figueiredo. "Bayesian image segmentation using hidden fields: Supervised, unsupervised, and semi-supervised formulations". In: *2016 24th European Signal Processing Conference (EUSIPCO)*. 2016, pp. 523–527 (cit. on pp. 17, 21).

[Bishop, 2006] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006 (cit. on pp. 8, 18, 27, 31, 57, 143, 147).

[Blei et al., 2017] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877 (cit. on pp. 9–10, 93).

[Bodenstedt et al., 2018] S. Bodenstedt, M. Allan, A. Agustinos, X. Du, L. Garcia-Peraza-Herrera, et al. *Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery*. 2018. arXiv: 1805.02475 [cs.CV] (cit. on p. 1).

[Böhning, 1992] D. Böhning. "Multinomial logistic regression algorithm". In: *Annals of the Institute of Statistical Mathematics* 44.1 (Mar. 1992), pp. 197–200 (cit. on pp. 28–29).

[Bonavita et al., 2020] I. Bonavita, X. Rafael-Palou, M. Ceresa, G. Piella, V. Ribas, et al. "Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline". In: *Computer Methods and Programs in Biomedicine* 185 (2020), p. 105172 (cit. on pp. 108–109, 115, 129).

[Bouman & Shapiro, 1994] C. A. Bouman and M. Shapiro. "A multiscale random field model for Bayesian image segmentation". In: *IEEE Transactions on Image Processing* 3.2 (1994), pp. 162–177 (cit. on p. 23).

[Boykov & Jolly, 2001] Y. Boykov and M. Jolly. "Interactive graph cuts for optimal boundary region segmentation of objects in N-D images". In: *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. Vol. 1. 2001, 105–112 vol.1 (cit. on pp. 16, 21).

[Calabrò et al., 2020] L. Calabrò, S. Peters, J.-C. Soria, A. M. Di Giacomo, F. Barlesi, et al. "Challenges in lung cancer therapy during the COVID-19 pandemic". In: *The Lancet Respiratory Medicine* 8.6 (June 2020), pp. 542–544 (cit. on p. 141).

[Causey et al., 2018] J. L. Causey, J. Zhang, S. Ma, B. Jiang, J. A. Qualls, et al. "Highly accurate model for prediction of lung nodule malignancy with CT scans". In: *Scientific Reports* 8.1 (June 2018), p. 9286 (cit. on p. 107).

[Chabrier et al., 2006] S. Chabrier, B. Emile, C. Rosenberger, and H. Laurent. "Unsupervised Performance Evaluation of Image Segmentation". In: *EURASIP Journal on Advances in Signal Processing* 2006.1 (Dec. 2006), p. 096306 (cit. on p. 48).

[Chan & Vese, 1999] T. Chan and L. Vese. "An Active Contour Model without Edges". In: *Scale-Space Theories in Computer Vision*. Ed. by M. Nielsen, P. Johansen, O. F. Olsen, and J. Weickert. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 141–151 (cit. on p. 16).

[Chen et al., 2020] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, et al. "Deep Learning for Cardiac Image Segmentation: A Review". In: *Frontiers in Cardiovascular Medicine* 7 (2020), p. 25 (cit. on p. 2).

[Cheng et al., 2016] L. Cheng, L.-Y. Lo, N. L. S. Tang, D. Wang, and K.-S. Leung. "CrossNorm: a novel normalization strategy for microarray data in cancers". In: *Scientific Reports* 6.1 (Jan. 2016), p. 18898 (cit. on p. 171).

[Clough et al., 2020] J. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, et al. "A Topological Loss Function for Deep-Learning based Image Segmentation using Persistent Homology". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), pp. 1–1 (cit. on p. 137).

[Commowick et al., 2012] O. Commowick, A. Akhondi-Asl, and S. K. Warfield. "Estimating A Reference Standard Segmentation With Spatially Varying Performance Parameters: Local MAP STAPLE". In: *IEEE Transactions on Medical Imaging* 31.8 (Aug. 2012), pp. 1593–1606 (cit. on p. 75).

[Commowick et al., 2018] O. Commowick, A. Istace, M. Kain, B. Laurent, et al. "Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure". In: *Scientific Reports* 8.1 (2018), p. 13650 (cit. on p. 57).

[Commowick & Warfield, 2009] O. Commowick and S. K. Warfield. "A Continuous STAPLE for Scalar, Vector, and Tensor Images: An Application to DTI Analysis". In: *IEEE Transactions on Medical Imaging* 28.6 (June 2009), pp. 838–846 (cit. on pp. 76, 102).

[Commowick & Warfield, 2010] O. Commowick and S. K. Warfield. "Incorporating Priors on Expert Performance Parameters for Segmentation Validation and Label Fusion: A Maximum a Posteriori STAPLE". In: *Proceedings of the 13th International Conference on Medical Image Computing and Computer-Assisted Intervention: Part III*. MICCAI'10. Beijing, China: Springer-Verlag, 2010, pp. 25–32 (cit. on p. 75).

[Cordts et al., 2016] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (cit. on p. 1).

[Couraud et al., 2012] S. Couraud, G. Zalcman, B. Milleron, F. Morin, and P.-J. Souquet. "Lung cancer in never smokers – A review". In: *European Journal of Cancer* 48.9 (June 2012), pp. 1299–1311 (cit. on p. 10).

[Cremers et al., 2007] D. Cremers, M. Rousson, and R. Deriche. "A Review of Statistical Approaches to Level Set Segmentation: Integrating Color, Texture, Motion and Shape". In: *International Journal of Computer Vision* 72.2 (Apr. 2007), pp. 195–215 (cit. on p. 1).

[Crowe et al., 2017] E. M. Crowe, W. Alderson, J. Rossiter, and C. Kent. "Expertise Affects Inter-Observer Agreement at Peripheral Locations within a Brain Tumor". In: *Frontiers in Psychology* 8 (2017), p. 1628 (cit. on p. 3).

[da Silva et al., 2018] G. L. F. da Silva, T. L. A. Valente, A. C. Silva, A. C. de Paiva, and M. Gattass. "Convolutional neural network-based PSO for lung nodule false positive reduction on CT images". In: *Computer Methods and Programs in Biomedicine* 162 (2018), pp. 109–118 (cit. on p. 107).

[Davis & Goadrich, 2006] J. Davis and M. Goadrich. "The Relationship between Precision-Recall and ROC Curves". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 233–240 (cit. on pp. 129, 172).

[Defossez et al., 2019] G. Defossez, S. L. Guyader-Peyrou, Z. Uhry, P. Grosclaude, M. Colonna, et al. *Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018. Volume 1 – Tumeurs solides*. Saint-Maurice (Fra): Santé publique France, 2019. 372 pp. (cit. on p. 106).

[DeVries & Taylor, 2018] T. DeVries and G. W. Taylor. *Leveraging Uncertainty Estimates for Predicting Segmentation Quality*. 2018. arXiv: 1807.00502 [cs.CV] (cit. on pp. 47, 50).

[Dong et al., 2020] G. Dong, Y. Zou, J. Jiao, Y. Liu, S. Liu, et al. "TexNet: Texture Loss Based Network for Gastric Antrum Segmentation in Ultrasound". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Cham: Springer International Publishing, 2020, pp. 138–145 (cit. on p. 137).

[DSB, 2017] *Data Science Bowl 2017 – Can you improve lung cancer detection?* 2017. URL: https://www.kaggle.com/c/data-science-bowl-2017/ (visited on Apr. 24, 2021) (cit. on p. 108).

[Duma et al., 2019] N. Duma, R. Santana-Davila, and J. R. Molina. "Non-Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment". In: *Mayo Clinic Proceedings* 94.8 (Aug. 2019), pp. 1623–1640 (cit. on pp. 10, 12).

[Durham & Adcock, 2015] A. L. Durham and I. M. Adcock. "The relationship between COPD and lung cancer". In: *Lung Cancer* 90.2 (Nov. 2015), pp. 121–127 (cit. on p. 109).

[Erasmus et al., 2000] J. J. Erasmus, J. E. Connolly, H. P. McAdams, and V. L. Roggli. "Solitary Pulmonary Nodules: Part I. Morphologic Evaluation for Differentiation of Benign and Malignant Lesions". In: *RadioGraphics* 20.1 (2000), pp. 43–58 (cit. on pp. 13, 107).

[Everingham et al., 2010] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. "The Pascal Visual Object Classes (VOC) Challenge". In: *International Journal of Computer Vision* 88.2 (June 2010), pp. 303–338 (cit. on p. 2).

[Fenster & Chiu, 2005] A. Fenster and B. Chiu. "Evaluation of Segmentation algorithms for Medical Imaging". In: *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. 2005, pp. 7186–7189 (cit. on pp. 96, 101).

[Ferdinand Christ et al., 2017] P. Ferdinand Christ, F. Ettlinger, G. Kaissis, S. Schlecht, F. Ahmaddy, et al. "SurvivalNet: Predicting patient survival from diffusion weighted magnetic resonance images using cascaded fully convolutional and 3D Convolutional Neural Networks". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. 2017, pp. 839–843 (cit. on p. 1).

[Figueiredo, 2005a] M. A. T. Figueiredo. "Bayesian image segmentation using wavelet-based priors". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 2005, 437–443 vol. 1 (cit. on p. 43).

[Figueiredo, 2005b] M. A. T. Figueiredo. "Bayesian Image Segmentation Using Gaussian Field Priors". In: *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Ed. by A. Rangarajan, B. Vemuri, and A. L. Yuille. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 74–89 (cit. on pp. 17, 19).

[Fortunato et al., 2019] O. Fortunato, P. Gasparini, M. Boeri, and G. Sozzi. "Exo-miRNAs as a New Tool for Liquid Biopsy in Lung Cancer". In: *Cancers* 11.6 (2019) (cit. on p. 170).

[Gal & Ghahramani, 2016] Y. Gal and Z. Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 1050–1059 (cit. on p. 16).

[Gao et al., 2017] H. Gao, Y. Tang, L. Jing, H. Li, and H. Ding. "A Novel Unsupervised Segmentation Quality Evaluation Method for Remote Sensing Images". In: *Sensors* 17.10 (Sept. 2017), p. 2427 (cit. on p. 48).

[Garcia-Garcia et al., 2017] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. *A Review on Deep Learning Techniques Applied to Semantic Segmentation*. 2017. arXiv: 1704.06857 [cs.CV] (cit. on p. 1).

[Ge et al., 2005] Z. Ge, B. Sahiner, H.-P. Chan, L. M. Hadjiiski, et al. "Computer-aided detection of lung nodules: False positive reduction using a 3D gradient field method and 3D ellipsoid fitting". In: *Medical Physics* 32.8 (2005), pp. 2443–2454 (cit. on p. 107).

[Genovesi et al., 2011] D. Genovesi, G. A. Cèfaro, A. Vinciguerra, A. Augurio, M. Di Tommaso, et al. "Interobserver variability of clinical target volume delineation in supra-diaphragmatic Hodgkin's disease". In: *Strahlentherapie und Onkologie* 187.6 (June 2011), pp. 357–366 (cit. on p. 2).

[Giri, 2016] R. Giri. "Bayesian sparse signal recovery using scale mixtures with applications to speech". PhD thesis. UC San Diego, 2016 (cit. on p. 79).

[Girshick et al., 2014] R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 580–587 (cit. on p. 118).

[Gómez et al., 1998] E. Gómez, M. Gomez-Viilegas, and J. Marín. "A multivariate generalization of the power exponential family of distributions". In: *Communications in Statistics - Theory and Methods* 27.3 (1998), pp. 589–600 (cit. on p. 79).

[Gómez-Sánchez-Manzano et al., 2008] E. Gómez-Sánchez-Manzano, M. A. Gómez-Villegas, and J. M. Marín. "Multivariate Exponential Power Distributions as Mixtures of Normal Distributions with Bayesian Applications". In: *Communications in Statistics - Theory and Methods* 37.6 (2008), pp. 972–985 (cit. on p. 80).

[Gould et al., 2007] M. K. Gould, J. Fletcher, M. D. Iannettoni, W. R. Lynch, D. E. Midthun, et al. "Evaluation of Patients With Pulmonary Nodules: When Is It Lung Cancer?: ACCP Evidence-Based Clinical Practice Guidelines (2nd Edition)". In: *CHEST* 132.3 (Sept. 2007), 108S–130S (cit. on p. 107).

[Greenspan et al., 2006] H. Greenspan, A. Ruf, and J. Goldberger. "Constrained Gaussian mixture model framework for automatic segmentation of MR brain images". In: *IEEE Transactions on Medical Imaging* 25.9 (2006), pp. 1233–1245 (cit. on p. 2).

[Gridelli et al., 2015] C. Gridelli, A. Rossi, D. P. Carbone, J. Guarize, N. Karachaliou, et al. "Non-small-cell lung cancer". In: *Nature Reviews Disease Primers* 1.1 (May 2015), p. 15009 (cit. on p. 10).

[Hancock & Magnan, 2016] M. C. Hancock and J. F. Magnan. “Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods”. In: *Journal of Medical Imaging* 3.4 (2016), pp. 1–15 (cit. on p. 110).

[He et al., 2016] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778 (cit. on p. 116).

[He & Garcia, 2009] H. He and E. A. Garcia. “Learning from Imbalanced Data”. In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284 (cit. on pp. 108, 129, 172).

[Heeke et al., 2019] S. Heeke, J. Benzaquen, E. Long-Mira, B. Audelan, V. Lespinet, et al. “In-house Implementation of Tumor Mutational Burden Testing to Predict Durable Clinical Benefit in Non-small Cell Lung Cancer and Melanoma Patients”. In: *Cancers* 11.9 (2019) (cit. on p. 5).

[Held et al., 1997] K. Held, E. R. Kops, B. J. Krause, W. M. Wells, R. Kikinis, et al. “Markov random field segmentation of brain MR images”. In: *IEEE Transactions on Medical Imaging* 16.6 (1997), pp. 878–886 (cit. on pp. 17, 20).

[Heller et al., 2018] N. Heller, J. Dean, and N. Papanikolopoulos. “Imperfect Segmentation Labels: How Much Do They Matter?” In: *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Cham: Springer International Publishing, 2018, pp. 112–120 (cit. on p. 2).

[Hoeting et al., 1999] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. “Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors”. In: *Statistical Science* 14.4 (1999), pp. 382–417 (cit. on p. 74).

[Huang et al., 2010] Z. Huang, D. Huang, S. Ni, Z. Peng, W. Sheng, et al. “Plasma microRNAs are promising novel biomarkers for early detection of colorectal cancer”. In: *International Journal of Cancer* 127.1 (2010), pp. 118–126 (cit. on p. 170).

[Huang et al., 2016] C. Huang, Q. Wu, and F. Meng. “QualityNet: Segmentation quality evaluation with deep convolutional networks”. In: *2016 Visual Communications and Image Processing (VCIP)*. Nov. 2016, pp. 1–4 (cit. on p. 47).

[Huang et al., 2017] Z. Huang, D. Zhu, L. Wu, M. He, X. Zhou, et al. “Six Serum-Based miRNAs as Potential Diagnostic Biomarkers for Gastric Cancer”. In: 26.2 (2017), pp. 188–196 (cit. on p. 170).

[Hui Zhang et al., 2006] Hui Zhang, S. Cholleti, S. A. Goldman, and J. E. Fritts. “Meta-Evaluation of Image Segmentation Using Machine Learning”. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*. Vol. 1. June 2006, pp. 1138–1145 (cit. on p. 47).

[Iglesias & Sabuncu, 2015] J. E. Iglesias and M. R. Sabuncu. “Multi-atlas segmentation of biomedical images: A survey”. In: *Medical Image Analysis* 24.1 (2015), pp. 205–219 (cit. on p. 1).

[Iqbal et al., 2019] M. A. Iqbal, S. Arora, G. Prakasam, G. A. Calin, and M. A. Syed. "MicroRNA in lung cancer: role, mechanisms, pathways and therapeutic relevance". In: *Molecular Aspects of Medicine* 70 (2019). Non-coding RNAs and DNAs in health and disease, pp. 3–20 (cit. on p. 170).

[Jaakkola & Jordan, 2000] T. S. Jaakkola and M. I. Jordan. "Bayesian parameter estimation via variational methods". In: *Statistics and Computing* 10.1 (Jan. 2000), pp. 25–37 (cit. on pp. 28, 56).

[Jacobs et al., 2016] C. Jacobs, E. M. van Rikxoort, K. Murphy, M. Prokop, C. M. Schaefer-Prokop, et al. "Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database". In: *European Radiology* 26.7 (July 2016), pp. 2139–2147 (cit. on p. 132).

[Jaeger et al., 2018] P. F. Jaeger, S. A. A. Kohl, S. Bickelhaupt, F. Isensee, T. A. Kuder, et al. *Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection*. 2018. arXiv: 1811.08661 [cs.CV] (cit. on pp. 107, 116).

[Jakab, 2012] A. Jakab. *Segmenting Brain Tumors with the Slicer 3D Software*. 2012. URL: http://www2.imm.dtu.dk/projects/BRATS2012/Jakab_TumorSegmentation_Manual.pdf (visited on June 22, 2020) (cit. on p. 61).

[Jeanneret-Sozzi et al., 2006] W. Jeanneret-Sozzi, R. Moeckli, J.-F. Valley, A. Zouhair, E. M. Ozsahin, et al. "The Reasons for Discrepancies in TargetVolume Delineation". In: *Strahlentherapie und Onkologie* 182.8 (Aug. 2006), pp. 450–457 (cit. on p. 2).

[Jin et al., 2018] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura. "CT-Realistic Lung Nodule Simulation from 3D Conditional Generative Adversarial Networks for Robust Lung Segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Ed. by A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger. Cham: Springer International Publishing, 2018, pp. 732–740 (cit. on p. 133).

[Johnson & Xie, 2011] B. Johnson and Z. Xie. "Unsupervised image segmentation evaluation and refinement using a multi-scale approach". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 66.4 (2011), pp. 473–483 (cit. on pp. 48, 58, 155).

[Joskowicz et al., 2019] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna. "Inter-observer variability of manual contour delineation of structures in CT". In: *European Radiology* 29.3 (Mar. 2019), pp. 1391–1399 (cit. on p. 74).

[Jungo & Reyes, 2019] A. Jungo and M. Reyes. "Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, et al. Cham: Springer International Publishing, 2019, pp. 48–56 (cit. on pp. 16, 47, 50).

[Kamnitsas et al., 2018] K. Kamnitsas, W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, et al. "Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation". In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham: Springer International Publishing, 2018, pp. 450–462 (cit. on p. 2).

[Kaseda, 2020] K. Kaseda. “Recent and Current Advances in FDG-PET Imaging within the Field of Clinical Oncology in NSCLC: A Review of the Literature”. In: *Diagnostics* 10.8 (2020) (cit. on p. 11).

[Keshavan et al., 2018] A. Keshavan, E. Datta, I. M. McDonough, C. R. Madan, K. Jordan, et al. “Mindcontrol: A web application for brain segmentation quality control”. In: *NeuroImage* 170 (2018). Segmenting the Brain, pp. 365–372 (cit. on p. 46).

[Kim et al., 2019] B.-C. Kim, J. S. Yoon, J.-S. Choi, and H.-I. Suk. “Multi-scale gradual integration CNN for false positive reduction in pulmonary nodule detection”. In: *Neural Networks* 115 (2019), pp. 1–10 (cit. on p. 119).

[Kocak et al., 2019] B. Kocak, E. S. Durmaz, O. K. Kaya, E. Ates, and O. Kilickesmez. “Reliability of Single-Slice–Based 2D CT Texture Analysis of Renal Masses: Influence of Intra- and Interobserver Manual Segmentation Variability on Radiomic Feature Reproducibility”. In: *American Journal of Roentgenology* 213.2 (Aug. 2019), pp. 377–383 (cit. on p. 74).

[Kohlberger et al., 2012] T. Kohlberger, V. Singh, C. Alvino, C. Bahlmann, et al. “Evaluating Segmentation Error without Ground Truth”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*. Ed. by N. Ayache, H. Delingette, P. Golland, and K. Mori. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 528–536 (cit. on pp. 47, 70).

[Kommrusch & Pouchet, 2019] S. Kommrusch and L. Pouchet. “Synthetic Lung Nodule 3D Image Generation Using Autoencoders”. In: *2019 International Joint Conference on Neural Networks (IJCNN)*. 2019, pp. 1–9 (cit. on p. 133).

[Koning et al., 2020] H. J. de Koning, C. M. van der Aalst, P. A. de Jong, E. T. Scholten, K. Nackaerts, et al. “Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial”. In: *New England Journal of Medicine* 382.6 (2020), pp. 503–513 (cit. on pp. 3, 106).

[Kozintsev, 1999] B. Kozintsev. “Computations with gaussian random fields”. PhD thesis. Institute for Systems Research, University of Maryland, 1999 (cit. on pp. 23, 30).

[Kristensen et al., 2017] I. Kristensen, K. Nilsson, M. Agrup, K. Belfrage, A. Embring, et al. “A dose based approach for evaluation of inter-observer variations in target delineation”. In: *Technical Innovations & Patient Support in Radiation Oncology* 3-4 (2017), pp. 41–47 (cit. on p. 2).

[Kumar et al., 2016] S. Kumar, G. Liney, R. Rai, L. Holloway, D. Moses, et al. “Magnetic resonance imaging in lung: a review of its potential for radiotherapy”. In: *The British Journal of Radiology* 89.1060 (2016), p. 20150431 (cit. on p. 11).

[Lampert et al., 2016] T. A. Lampert, A. Stumpf, and P. Gançarski. “An Empirical Study Into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation”. In: *IEEE Transactions on Image Processing* 25.6 (2016), pp. 2557–2572 (cit. on pp. 74, 101).

[Landman et al., 2012] B. A. Landman, A. J. Asman, A. G. Scoggins, J. A. Bogovic, F. Xing, et al. “Robust Statistical Fusion of Image Labels”. In: *IEEE Transactions on Medical Imaging* 31.2 (2012), pp. 512–522 (cit. on p. 75).

[Langerak et al., 2010] T. R. Langerak, U. A. van der Heide, A. N. T. J. Kotte, et al. "Label Fusion in Atlas-Based Segmentation Using a Selective and Iterative Method for Performance Level Estimation (SIMPLE)". In: *IEEE Transactions on Medical Imaging* 29.12 (Dec. 2010), pp. 2000–2008 (cit. on p. 76).

[Langlotz et al., 2019] C. P. Langlotz, B. Allen, B. J. Erickson, J. Kalpathy-Cramer, K. Bigelow, et al. "A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop". In: *Radiology* 291.3 (2019), pp. 781–791 (cit. on p. 2).

[Le Folgoc et al., 2017] L. Le Folgoc, H. Delingette, A. Criminisi, and N. Ayache. "Sparse Bayesian registration of medical images for self-tuning of parameters and spatially adaptive parametrization of displacements". In: *Medical Image Analysis* 36 (2017), pp. 79–97 (cit. on p. 24).

[LeCun et al., 2015] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". In: *Nature* 521.7553 (May 2015), pp. 436–444 (cit. on p. 2).

[Leroy et al., 2017] S. Leroy, J. Benzaquen, A. Mazzetta, S. Marchand-Adam, B. Padovani, et al. "Circulating tumour cells as a potential screening tool for lung cancer (the AIR study): protocol of a prospective multicentre cohort study in France". In: *BMJ Open* 7.12 (2017) (cit. on p. 109).

[Li et al., 2009] X. A. Li, A. Tai, D. W. Arthur, T. A. Buchholz, S. Macdonald, et al. "Variability of Target and Normal Structure Delineation for Breast Cancer Radiotherapy: An RTOG Multi-Institutional and Multiobserver Study". In: *International Journal of Radiation Oncology*Biology*Physics* 73.3 (2009), pp. 944–951 (cit. on p. 3).

[Li et al., 2015] C. Li, X. Wang, S. Eberl, M. Fulham, Y. Yin, et al. "Supervised Variational Model With Statistical Inference and Its Application in Medical Image Segmentation". In: *IEEE Transactions on Biomedical Engineering* 62.1 (2015), pp. 196–207 (cit. on p. 18).

[Liao et al., 2019] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song. "Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-OR Network". In: *IEEE Transactions on Neural Networks and Learning Systems* 30.11 (2019), pp. 3484–3495 (cit. on pp. 4, 108–109, 116, 119, 129).

[Lin et al., 2014] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, et al. "Microsoft COCO: Common Objects in Context". In: *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014, pp. 740–755 (cit. on pp. 2, 57, 138).

[Litjens et al., 2014] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman. "Computer-Aided Detection of Prostate Cancer in MRI". In: *IEEE Transactions on Medical Imaging* 33.5 (2014), pp. 1083–1092 (cit. on p. 88).

[Litjens et al., 2017] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88 (cit. on p. 16).

[Liu et al., 2010] F. Liu, B. Zhao, L. M. Krug, N. M. Ishill, R. C. Lim, et al. "Assessment of Therapy Responses and Prediction of Survival in Malignant Pleural Mesothelioma Through Computer-Aided Volumetric Measurement on Computed Tomography Scans". In: *Journal of Thoracic Oncology* 5.6 (2010), pp. 879–884 (cit. on p. 1).

[Liu et al., 2013] X. Liu, A. Montillo, E. T. Tan, and J. F. Schenck. "iSTAPLE: improved label fusion for segmentation by combining STAPLE with image intensity". In: *Medical Imaging 2013: Image Processing*. Ed. by S. Ourselin and D. R. Haynor. Vol. 8669. International Society for Optics and Photonics. SPIE, 2013, pp. 727–732 (cit. on p. 75).

[Liu et al., 2020] J. Liu, X. Wang, and X.-c. Tai. *Deep Convolutional Neural Networks with Spatial Regularization, Volume and Star-shape Priori for Image Segmentation*. 2020. arXiv: 2002.03989 [cs.CV] (cit. on p. 137).

[Louie et al., 2010] A. V. Louie, G. Rodrigues, J. Olsthoorn, D. Palma, E. Yu, et al. "Inter-observer and intra-observer reliability for lung cancer target volume delineation in the 4D-CT era". In: *Radiotherapy and Oncology* 95.2 (2010), pp. 166–171 (cit. on p. 2).

[Luo et al., 2020] J. Luo, H. Rizvi, I. R. Preeshagul, J. V. Egger, D. Hoyos, et al. "COVID-19 in patients with lung cancer". In: *Annals of Oncology* 31.10 (Sept. 2020), pp. 1386–1396 (cit. on p. 141).

[MacKay, 1999] D. J. C. MacKay. "Comparison of Approximate Methods for Handling Hyperparameters". In: *Neural Computation* 11.5 (1999), pp. 1035–1068 (cit. on p. 8).

[MacMahon et al., 2005] H. MacMahon, J. H. M. Austin, G. Gamsu, C. J. Herold, J. R. Jett, et al. "Guidelines for Management of Small Pulmonary Nodules Detected on CT Scans: A Statement from the Fleischner Society". In: *Radiology* 237.2 (2005), pp. 395–400 (cit. on p. 13).

[Madhavan et al., 2016] D. Madhavan, C. Peng, M. Wallwiener, M. Zucknick, J. Nees, et al. "Circulating miRNAs with prognostic value in metastatic breast cancer and for early detection of metastasis". In: *Carcinogenesis* 37.5 (Jan. 2016), pp. 461–470 (cit. on p. 170).

[Malsiner-Walli et al., 2017] G. Malsiner-Walli, S. Frühwirth-Schnatter, and B. Grün. "Identifying Mixtures of Mixtures Using Bayesian Estimation". In: *Journal of Computational and Graphical Statistics* 26.2 (2017), pp. 285–295 (cit. on p. 18).

[Mazzara et al., 2004] G. P. Mazzara, R. P. Velthuizen, J. L. Pearlman, H. M. Greenberg, and H. Wagner. "Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation". In: *International Journal of Radiation Oncology*Biology*Physics* 59.1 (2004), pp. 300–312 (cit. on p. 1).

[McCunney & Li, 2014] R. J. McCunney and J. Li. "Radiation Risks in Lung Cancer Screening Programs". In: *Chest* 145.3 (2014), pp. 618–624 (cit. on p. 169).

[McDonald & Newey, 1988] J. B. McDonald and W. K. Newey. "Partially Adaptive Estimation of Regression Models via the Generalized t Distribution". In: *Econometric Theory* 4.3 (1988), pp. 428–457 (cit. on p. 79).

[McNitt-Gray et al., 2007] M. F. McNitt-Gray, S. G. Armato III, C. R. Meyer, A. P. Reeves, G. McLennan, et al. "The Lung Image Database Consortium (LIDC) Data Collection Process for Nodule Detection and Annotation". In: *Academic Radiology* 14.12 (Dec. 2007), pp. 1464–1474 (cit. on pp. 110, 130).

[Meng et al., 2017] Q. Meng, D. Catchpoole, D. Skillicorn, and P. J. Kennedy. "DBNorm: normalizing high-density oligonucleotide microarray data based on distributions". In: *BMC Bioinformatics* 18.1 (Nov. 2017), p. 527 (cit. on p. 171).

[Menze et al., 2015] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)". In: *IEEE Transactions on Medical Imaging* 34.10 (Sept. 2015), pp. 1993–2024 (cit. on pp. 3, 57, 62, 74).

[Menze et al., 2020] B. Menze, L. Joskowicz, S. Bakas, A. Jakab, E. Konukoglu, et al. *Quantification of Uncertainties in Biomedical Image Quantification Challenge*. 2020. URL: https://qubiq.grand-challenge.org/Home/ (visited on Jan. 6, 2021) (cit. on p. 34).

[Meyer et al., 2019] A. Meyer, M. Rakr, D. Schindele, S. Blaschke, M. Schostak, et al. "Towards Patient-Individual PI-Rads v2 Sector Map: Cnn for Automatic Segmentation of Prostatic Zones From T2-Weighted MRI". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019, pp. 696–700 (cit. on p. 88).

[Miller et al., 2017] A. C. Miller, N. J. Foti, and R. P. Adams. "Variational Boosting: Iteratively Refining Posterior Approximations". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, Aug. 2017, pp. 2420–2429 (cit. on pp. 76, 86).

[Mlynarski et al., 2019] P. Mlynarski, H. Delingette, A. Criminisi, and N. Ayache. "3D convolutional neural networks for tumor segmentation using long-range 2D context". In: *Computerized Medical Imaging and Graphics* 73 (2019), pp. 60–72 (cit. on p. 69).

[Montani et al., 2015] F. Montani, M. J. Marzi, F. Dezi, E. Dama, R. M. Carletti, et al. "miR-Test: A Blood Test for Lung Cancer Early Detection". In: *JNCI: Journal of the National Cancer Institute* 107.6 (Mar. 2015) (cit. on pp. 170, 172).

[Morris et al., 1997] R. Morris, X. Descombes, and J. Zerubia. "Fully Bayesian Image Segmentation-an Engineering Perspective". In: *Proceedings of the 1997 International Conference on Image Processing (ICIP '97) 3-Volume Set-Volume 3 - Volume 3*. ICIP '97. USA: IEEE Computer Society, 1997, p. 54 (cit. on p. 38).

[Murphy, 2012] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012 (cit. on pp. 18, 28–29).

[Nair et al., 2020] T. Nair, D. Precup, D. L. Arnold, and T. Arbel. "Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation". In: *Medical Image Analysis* 59 (2020), p. 101557 (cit. on p. 16).

[Nakamura et al., 2008] K. Nakamura, Y. Shioyama, S. Tokumaru, N. Hayashi, N. Oya, et al. "Variation of Clinical Target Volume Definition among Japanese Radiation Oncologists in External Beam Radiotherapy for Prostate Cancer". In: *Japanese Journal of Clinical Oncology* 38.4 (Mar. 2008), pp. 275–280 (cit. on p. 2).

[Naqi et al., 2018] S. M. Naqi, M. Sharif, and M. Yasmin. "Multistage segmentation model and SVM-ensemble for precise lung nodule detection". In: *International Journal of Computer Assisted Radiology and Surgery* 13.7 (July 2018), pp. 1083–1095 (cit. on p. 107).

[Neubauer et al., 2016] J. Neubauer, E. M. Spira, J. Strube, M. Langer, C. Voss, et al. "Image quality of mixed convolution kernel in thoracic computed tomography". In: *Medicine* 95.44 (2016) (cit. on p. 132).

[Neville, 2013] S. E. Neville. "Elaborate distribution semiparametric regression via mean field variational Bayes". PhD thesis. University of Wollongong, 2013 (cit. on p. 162).

[NLST, 2011] The National Lung Screening Trial Research Team. "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening". In: *New England Journal of Medicine* 365.5 (2011), pp. 395–409 (cit. on pp. 3, 106, 108, 130).

[Nosrati & Hamarneh, 2016] M. S. Nosrati and G. Hamarneh. *Incorporating prior knowledge in medical image segmentation: a survey*. 2016. arXiv: 1607.01092 [cs.CV] (cit. on pp. 7, 16–17).

[Orbanz & Buhmann, 2005] P. Orbanz and J. M. Buhmann. "SAR images as mixtures of Gaussian mixtures". In: *IEEE International Conference on Image Processing 2005*. Vol. 2. 2005, pp. II–209 (cit. on p. 18).

[Orbanz & Buhmann, 2008] P. Orbanz and J. M. Buhmann. "Nonparametric Bayesian Image Segmentation". In: *International Journal of Computer Vision* 77.1 (May 2008), pp. 25–45 (cit. on p. 43).

[Ost et al., 2003] D. Ost, A. M. Fein, and S. H. Feinsilver. "The Solitary Pulmonary Nodule". In: *New England Journal of Medicine* 348.25 (2003). PMID: 12815140, pp. 2535–2542 (cit. on pp. 106–107).

[Ozdemir et al., 2020] O. Ozdemir, R. L. Russell, and A. A. Berlin. "A 3D Probabilistic Deep Learning System for Detection and Diagnosis of Lung Cancer Using Low-Dose CT Scans". In: *IEEE Transactions on Medical Imaging* 39.5 (2020), pp. 1419–1429 (cit. on pp. 108–109, 115, 129).

[Papadopoulos et al., 2016] D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari. "We Don't Need No Bounding-Boxes: Training Object Class Detectors Using Only Human Verification". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 854–863 (cit. on p. 138).

[Park & Casella, 2008] T. Park and G. Casella. "The Bayesian Lasso". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 681–686 (cit. on p. 75).

[Pascal et al., 2013] F. Pascal, L. Bombrun, J. Tourneret, and Y. Berthoumieu. "Parameter Estimation For Multivariate Generalized Gaussian Distributions". In: *IEEE Transactions on Signal Processing* 61.23 (2013), pp. 5960–5971 (cit. on p. 79).

[Pastorino et al., 2019] U. Pastorino, N. Sverzellati, S. Sestini, M. Silva, F. Sabia, et al. "Ten-year results of the Multicentric Italian Lung Detection trial demonstrate the safety and efficacy of biennial lung cancer screening". In: *European Journal of Cancer* 118 (Sept. 2019), pp. 142–148 (cit. on p. 106).

[Patz et al., 2014] J. Patz Edward F., P. Pinsky, C. Gatsonis, J. D. Sicks, B. S. Kramer, et al. "Overdiagnosis in Low-Dose Computed Tomography Screening for Lung Cancer". In: *JAMA Internal Medicine* 174.2 (Feb. 2014), pp. 269–274 (cit. on p. 130).

[Pehrson et al., 2019] L. M. Pehrson, M. B. Nielsen, and C. Ammitzbøl Lauridsen. "Automatic Pulmonary Nodule Detection Applying Deep Learning or Machine Learning Algorithms to the LIDC-IDRI Database: A Systematic Review". In: *Diagnostics* 9.1 (2019) (cit. on p. 107).

[Pereyra et al., 2015] M. Pereyra, J. M. Bioucas-Dias, and M. A. T. Figueiredo. "Maximum-a-posteriori estimation with unknown regularisation parameters". In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. 2015, pp. 230–234 (cit. on pp. 22, 53).

[Pereyra & McLaughlin, 2017] M. Pereyra and S. McLaughlin. “Fast Unsupervised Bayesian Image Segmentation With Adaptive Spatial Regularisation”. In: *IEEE Transactions on Image Processing* 26.6 (June 2017), pp. 2577–2587 (cit. on p. 21).

[Petersen et al., 2007] R. P. Petersen, P. T. Truong, H. A. Kader, E. Berthelet, J. C. Lee, et al. “Target Volume Delineation for Partial Breast Radiotherapy Planning: Clinical Characteristics Associated with Low Interobserver Concordance”. In: *International Journal of Radiation Oncology*Biology*Physics* 69.1 (2007), pp. 41–48 (cit. on p. 74).

[Pohl et al., 2007] K. M. Pohl, J. Fisher, S. Bouix, M. Shenton, R. W. McCarley, et al. “Using the logarithm of odds to define a vector space on probabilistic atlases”. In: *Medical Image Analysis* 11.5 (2007). Special Issue on the Ninth International Conference on Medical Image Computing and Computer-Assisted Interventions - MICCAI 2006, pp. 465–477 (cit. on pp. 76–77).

[Rampinelli et al., 2013] C. Rampinelli, D. Origgi, and M. Bellomi. “Low-dose CT: technique, reading methods and image interpretation”. In: *Cancer imaging : the official publication of the International Cancer Imaging Society* 12.3 (Feb. 2013), pp. 548–556 (cit. on p. 12).

[Rasmussen & Williams, 2005] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005 (cit. on pp. 22–24).

[Ren et al., 2017] S. Ren, K. He, R. Girshick, and J. Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1137–1149 (cit. on p. 117).

[Rikxoort et al., 2009] E. M. van Rikxoort, B. de Hoop, M. A. Viergever, M. Prokop, and B. van Ginneken. “Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection”. In: *Medical Physics* 36.7 (2009), pp. 2934–2947 (cit. on p. 115).

[Robinson et al., 2018] R. Robinson, O. Oktay, W. Bai, V. V. Valindria, et al. “Real-Time Prediction of Segmentation Quality”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Cham: Springer International Publishing, 2018, pp. 578–585 (cit. on pp. 47, 70).

[Robinson et al., 2019] R. Robinson, V. V. Valindria, W. Bai, O. Oktay, et al. “Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study”. In: *Journal of Cardiovascular Magnetic Resonance* 21.1 (2019), p. 18 (cit. on p. 47).

[Roche et al., 2011] A. Roche, D. Ribes, M. Bach-Cuadra, and G. Krüger. “On the convergence of EM-like algorithms for image segmentation using Markov random fields”. In: *Medical Image Analysis* 15.6 (2011), pp. 830–839 (cit. on p. 55).

[Ronneberger et al., 2015] O. Ronneberger, P. Fischer, and T. Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241 (cit. on pp. 88, 115).

[Rosenberger et al., 2006] C. Rosenberger, S. Chabrier, H. Laurent, and B. Emile. "Unsupervised and supervised image segmentation evaluation". In: *Advances in Image and Video Segmentation, IGI Global* (Jan. 2006), pp. 365–393 (cit. on p. 48).

[Rother et al., 2004] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut -Interactive Foreground Extraction using Iterated Graph Cuts". In: *ACM Transactions on Graphics (SIGGRAPH)* (Aug. 2004) (cit. on pp. 16–17, 21, 52).

[Roy et al., 2019] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger. "Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control". In: *NeuroImage* 195 (2019), pp. 11–22 (cit. on pp. 47, 50).

[Rubin, 2015] G. D. Rubin. "Lung Nodule and Cancer Detection in Computed Tomography Screening". In: *Journal of Thoracic Imaging* 30.2 (2015) (cit. on pp. 106–107).

[Saatchi, 2011] Y. Saatchi. "Scalable Inference for Structured Gaussian Process Models". PhD thesis. University of Cambridge, 2011 (cit. on pp. 23, 30).

[Sabuncu et al., 2010] M. R. Sabuncu, B. T. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland. "A Generative Model for Image Segmentation Based on Label Fusion". In: *IEEE Transactions on Medical Imaging* 29.10 (2010), pp. 1714–1729 (cit. on p. 74).

[Sabuncu & Van Leemput, 2012] M. R. Sabuncu and K. Van Leemput. "The Relevance Voxel Machine (RVoxM): A Self-Tuning Bayesian Model for Informative Image-Based Prediction". In: *IEEE Transactions on Medical Imaging* 31.12 (2012), pp. 2290–2306 (cit. on p. 31).

[Sadeghigol et al., 2016] Z. Sadeghigol, M. H. Kahaei, and F. Haddadi. "Model based variational Bayesian compressive sensing using heavy tailed sparse prior". In: *Signal Processing: Image Communication* 41 (2016), pp. 158–167 (cit. on p. 76).

[Salimans et al., 2015] T. Salimans, D. Kingma, and M. Welling. "Markov Chain Monte Carlo and Variational Inference: Bridging the Gap". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 1218–1226 (cit. on p. 8).

[Schwarzenbach et al., 2015] H. Schwarzenbach, A. M. da Silva, G. Calin, and K. Pantel. "Data Normalization Strategies for MicroRNA Quantification". In: *Clinical Chemistry* 61.11 (Nov. 2015), pp. 1333–1342 (cit. on p. 171).

[Seeram, 2016] E. Seeram. *Computed tomography : physical principles, clinical applications, and quality control*. English. 2016 (cit. on p. 132).

[Setio et al., 2016] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, et al. "Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks". In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1160–1169 (cit. on p. 107).

[Setio et al., 2017] A. A. A. Setio, A. Traverso, T. de Bel, M. S. Berens, C. van den Bogaard, et al. "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge". In: *Medical Image Analysis* 42 (2017), pp. 1–13 (cit. on pp. 107–108, 110, 123, 138).

[Shamir & Bomzon, 2019] R. R. Shamir and Z. Bomzon. *Evaluation of head segmentation quality for treatment planning of tumor treating fields in brain tumors*. 2019. arXiv: 1906.11014 [eess.IV] (cit. on p. 47).

[Shaukat et al., 2017] F. Shaukat, G. Raja, A. Gooya, and A. F. Frangi. "Fully automatic detection of lung nodules in CT images using a hybrid feature set". In: *Medical Physics* 44.7 (2017), pp. 3615–3629 (cit. on p. 107).

[Shen et al., 2011] J. Shen, Z. Liu, N. W. Todd, H. Zhang, J. Liao, et al. "Diagnosis of lung cancer in individuals with solitary pulmonary nodules by plasma microRNA biomarkers". In: *BMC Cancer* 11.1 (Aug. 2011), p. 374 (cit. on p. 170).

[Shi et al., 2017] W. Shi, F. Meng, and Q. Wu. "Segmentation quality evaluation based on multi-scale convolutional neural networks". In: *2017 IEEE Visual Communications and Image Processing (VCIP)*. Dec. 2017, pp. 1–4 (cit. on p. 47).

[Shi et al., 2021] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, et al. "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19". In: *IEEE Reviews in Biomedical Engineering* 14 (2021), pp. 4–15 (cit. on p. 1).

[Siegel et al., 2020] R. L. Siegel, K. D. Miller, and A. Jemal. "Cancer statistics, 2020". In: *CA: A Cancer Journal for Clinicians* 70.1 (2020), pp. 7–30 (cit. on p. 106).

[Snoeckx et al., 2018] A. Snoeckx, P. Reyntiens, D. Desbuquoit, M. J. Spinhoven, P. E. Van Schil, et al. "Evaluation of the solitary pulmonary nodule: size matters, but do not ignore the power of morphology". In: *Insights into Imaging* 9.1 (Feb. 2018), pp. 73–86 (cit. on pp. 12–13, 107).

[Sozzi et al., 2014] G. Sozzi, M. Boeri, M. Rossi, C. Verri, P. Suatoni, et al. "Clinical Utility of a Plasma-Based miRNA Signature Classifier Within Computed Tomography Lung Cancer Screening: A Correlative MILD Trial Study". In: *Journal of Clinical Oncology* 32.8 (2014), pp. 768–773 (cit. on p. 170).

[Stoehr, 2017] J. Stoehr. *A review on statistical inference methods for discrete Markov random fields*. 2017. arXiv: 1704.03331 [stat.ME] (cit. on p. 39).

[Su et al., 2012] H. Su, J. Deng, and L. Fei-Fei. "Crowdsourcing Annotations for Visual Object Detection". In: *The 4th Human Computation Workshop*. AAAI Workshops. AAAI Press, 2012 (cit. on p. 138).

[Sun et al., 2019] L. Sun, S. Zhang, H. Chen, and L. Luo. "Brain Tumor Segmentation and Survival Prediction Using Multimodal MRI Scans With Deep Learning". In: *Frontiers in Neuroscience* 13 (2019), p. 810 (cit. on p. 1).

[Taha & Hanbury, 2015] A. A. Taha and A. Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool". In: *BMC Medical Imaging* 15.1 (Aug. 2015), p. 29 (cit. on p. 101).

[Tang et al., 2021] Y. Tang, R. Gao, H. H. Lee, S. Han, Y. Chen, et al. "High-resolution 3D abdominal segmentation with random patch network fusion". In: *Medical Image Analysis* 69 (2021), p. 101894 (cit. on p. 74).

[Thurlbeck & Müller, 1994] W. M. Thurlbeck and N. L. Müller. "Emphysema: definition, imaging, and quantification." In: *American Journal of Roentgenology* 163.5 (Nov. 1994), pp. 1017–1025 (cit. on p. 132).

[Tipping, 2001] M. E. Tipping. “Sparse Bayesian Learning and the Relevance Vector Machine”. In: *J. Mach. Learn. Res.* 1 (Sept. 2001), pp. 211–244 (cit. on pp. 24, 31).

[Tipping & Faul, 2003] M. E. Tipping and A. C. Faul. “Fast Marginal Likelihood Maximisation for Sparse Bayesian Models”. In: *AISTATS*. 2003 (cit. on pp. 18, 24, 31–32, 151–152, 172).

[Torre et al., 2016] L. A. Torre, R. L. Siegel, and A. Jemal. “Lung Cancer Statistics”. In: *Lung Cancer and Personalized Medicine: Current Knowledge and Therapies*. Ed. by A. Ahmad and S. Gadgeel. Cham: Springer International Publishing, 2016, pp. 1–19 (cit. on p. 106).

[Tsim et al., 2010] S. Tsim, C. A. O’Dowd, R. Milroy, and S. Davidson. “Staging of non-small cell lung cancer (NSCLC): A review”. In: *Respiratory Medicine* 104.12 (Dec. 2010), pp. 1767–1774 (cit. on pp. 11–12).

[Unkelbach et al., 2014] J. Unkelbach, B. H. Menze, E. Konukoglu, F. Dittmann, M. Le, et al. “Radiotherapy planning for glioblastoma based on a tumor growth model: improving target volume delineation”. In: *Physics in Medicine and Biology* 59.3 (Jan. 2014), pp. 747–770 (cit. on p. 2).

[USPSTF, 2021] US Preventive Services Task Force. “Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement”. In: *JAMA* 325.10 (Mar. 2021), pp. 962–970 (cit. on pp. 12, 130).

[Valindria et al., 2017] V. V. Valindria, I. Lavdas, W. Bai, K. Kamnitsas, et al. “Reverse Classification Accuracy: Predicting Segmentation Performance in the Absence of Ground Truth”. In: *IEEE Transactions on Medical Imaging* 36.8 (Aug. 2017), pp. 1597–1606 (cit. on p. 47).

[Van Haren et al., 2021] R. M. Van Haren, A. M. Delman, K. M. Turner, B. Waits, M. Hemingway, et al. “Impact of the COVID-19 Pandemic on Lung Cancer Screening Program and Subsequent Lung Cancer”. In: *Journal of the American College of Surgeons* 232.4 (2021), pp. 600–605 (cit. on p. 141).

[Visser et al., 2019] M. Visser, D. Müller, R. van Duijn, M. Smits, N. Verburg, et al. “Inter-rater agreement in glioma segmentations on longitudinal MRI”. In: *NeuroImage: Clinical* 22 (2019), p. 101727 (cit. on p. 46).

[Wahidi et al., 2007] M. M. Wahidi, J. A. Govert, R. K. Goudar, M. K. Gould, and D. C. McCrory. “Evidence for the Treatment of Patients With Pulmonary Nodules: When Is It Lung Cancer?: ACCP Evidence-Based Clinical Practice Guidelines (2nd Edition)”. In: *CHEST* 132.3 (Sept. 2007), 94S–107S (cit. on p. 106).

[Wang et al., 2013] Y. Wang, D. Wang, G. Zhang, and J. Wang. “Estimating nitrogen status of rice using the image segmentation of G-R thresholding method”. In: *Field Crops Research* 149 (2013), pp. 33–39 (cit. on p. 1).

[Wang et al., 2019] Y. Wang, Y. Zhou, W. Shen, S. Park, E. K. Fishman, et al. “Abdominal multi-organ segmentation with organ-attention networks and statistical fusion”. In: *Medical Image Analysis* 55 (2019), pp. 88–102 (cit. on p. 74).

[Wang & Yeung, 2016] H. Wang and D. Yeung. “Towards Bayesian Deep Learning: A Framework and Some Existing Methods”. In: *IEEE Transactions on Knowledge and Data Engineering* 28.12 (2016), pp. 3395–3408 (cit. on p. 17).

[Warfield et al., 2002] S. K. Warfield, K. H. Zou, and W. M. Wells. "Validation of Image Segmentation and Expert Quality with an Expectation-Maximization Algorithm". In: *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2002*. Ed. by T. Dohi and R. Kikinis. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 298–306 (cit. on p. 93).

[Warfield et al., 2004] S. K. Warfield, K. H. Zou, and W. M. Wells. "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation". In: *IEEE Transactions on Medical Imaging* 23.7 (2004), pp. 903–921 (cit. on pp. 3, 17, 20, 75, 78, 91).

[Warfield et al., 2008] S. K. Warfield, K. H. Zou, and W. M. Wells. "Validation of image segmentation by estimating rater bias and variance". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 366.1874 (2008), pp. 2361–2375 (cit. on pp. 75, 77–78, 93).

[Willemink et al., 2020] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, et al. "Preparing Medical Imaging Data for Machine Learning". In: *Radiology* 295.1 (2020), pp. 4–15 (cit. on p. 2).

[Winkels & Cohen, 2019] M. Winkels and T. S. Cohen. "Pulmonary nodule detection in CT scans with equivariant CNNs". In: *Medical Image Analysis* 55 (2019), pp. 15–26 (cit. on p. 107).

[Wood et al., 2018] D. E. Wood, E. A. Kazerooni, S. L. Baum, G. A. Eapen, D. S. Ettinger, et al. "Lung Cancer Screening, Version 3.2018, NCCN Clinical Practice Guidelines in Oncology". In: *Journal of the National Comprehensive Cancer Network J Natl Compr Canc Netw* 16.4 (2018), pp. 412–441 (cit. on pp. 4, 106, 169).

[Woolrich et al., 2005] M. W. Woolrich, T. E. J. Behrens, C. F. Beckmann, and S. M. Smith. "Mixture models with adaptive spatial regularization for segmentation with an application to FMRI data". In: *IEEE Transactions on Medical Imaging* 24.1 (2005), pp. 1–11 (cit. on p. 21).

[Woolrich et al., 2009] M. W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, et al. "Bayesian analysis of neuroimaging data in FSL". In: *NeuroImage* 45.1, Supplement 1 (2009). Mathematics in Brain Imaging, S173–S186 (cit. on p. 20).

[Woolrich & Behrens, 2006] M. W. Woolrich and T. E. Behrens. "Variational bayes inference of spatial mixture models for segmentation". In: *IEEE Transactions on Medical Imaging* 25.10 (2006), pp. 1380–1391 (cit. on p. 21).

[Wu et al., 2018] B. Wu, Z. Zhou, J. Wang, and Y. Wang. "Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, pp. 1109–1113 (cit. on p. 107).

[Wu et al., 2020] J. Wu, Y. Gur, A. Karargyris, A. B. Syed, O. Boyko, et al. "Automatic Bounding Box Annotation of Chest X-Ray Data for Localization of Abnormalities". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 2020, pp. 799–803 (cit. on p. 138).

[Xie et al., 2018] Y. Xie, J. Zhang, Y. Xia, M. Fulham, and Y. Zhang. "Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT". In: *Information Fusion* 42 (2018), pp. 102–110 (cit. on p. 107).

[Xie et al., 2019a] H. Xie, D. Yang, N. Sun, Z. Chen, and Y. Zhang. "Automated pulmonary nodule detection in CT images using deep convolutional neural networks". In: *Pattern Recognition* 85 (2019), pp. 109–119 (cit. on p. 107).

[Xie et al., 2019b] Y. Xie, Y. Xia, J. Zhang, Y. Song, D. Feng, et al. "Knowledge-based Collaborative Deep Learning for Benign-Malignant Lung Nodule Classification on Chest CT". In: *IEEE Transactions on Medical Imaging* 38.4 (2019), pp. 991–1004 (cit. on p. 107).

[Xing et al., 2016] F. Xing, J. L. Prince, and B. A. Landman. "Investigation of Bias in Continuous Medical Image Label Fusion". In: *PLOS ONE* 11.6 (June 2016), pp. 1–15 (cit. on pp. 75, 77–78).

[Xu et al., 2009] Y. Xu, P. Kavanagh, M. Fish, J. Gerlach, et al. "Automated Quality Control for Segmentation of Myocardial Perfusion SPECT". In: *Journal of Nuclear Medicine* 50.9 (2009), pp. 1418–1426 (cit. on pp. 47, 70).

[Xu et al., 2010] J. Xu, J. P. Monaco, and A. Madabhushi. "Markov Random Field driven Region-Based Active Contour Model (MaRACel): Application to Medical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*. Ed. by T. Jiang, N. Navab, J. P. W. Pluim, and M. A. Viergever. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 197–204 (cit. on pp. 17, 20).

[Xu et al., 2019] Z. Xu, X. Wang, H.-C. Shin, H. Roth, D. Yang, et al. "Tunable CT Lung Nodule Synthesis Conditioned on Background Image and Semantic Features". In: *Simulation and Synthesis in Medical Imaging*. Cham: Springer International Publishing, 2019, pp. 62–70 (cit. on p. 133).

[Yang et al., 2019] J. Yang, S. Liu, S. Grbic, A. A. A. Setio, Z. Xu, et al. "Class-Aware Adversarial Lung Nodule Synthesis In CT Images". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019, pp. 1348–1352 (cit. on p. 133).

[Zhang et al., 2008] H. Zhang, J. E. Fritts, and S. A. Goldman. "Image segmentation evaluation: A survey of unsupervised methods". In: *Computer Vision and Image Understanding* 110.2 (2008), pp. 260–280 (cit. on pp. 46, 48, 58, 70, 153–155).

[Zhang et al., 2019] C. Zhang, X. Sun, K. Dang, K. Li, X.-w. Guo, et al. "Toward an Expert Level of Lung Cancer Detection and Classification Using a Deep Convolutional Neural Network". In: *The Oncologist* 24.9 (2019), pp. 1159–1165 (cit. on pp. 108, 129).

[Zhang et al., 2020] L. Zhang, R. Tanno, K. Bronik, C. Jin, P. Nachev, et al. "Learning to Segment When Experts Disagree". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Cham: Springer International Publishing, 2020, pp. 179–190 (cit. on p. 140).

[Zhao et al., 2019] Z. Zhao, P. Zheng, S. Xu, and X. Wu. "Object Detection With Deep Learning: A Review". In: *IEEE Transactions on Neural Networks and Learning Systems* 30.11 (2019), pp. 3212–3232 (cit. on p. 116).

[Zheng, 2016] M. Zheng. "Classification and Pathology of Lung Cancer". In: *Surgical Oncology Clinics* 25.3 (July 2016), pp. 447–468 (cit. on p. 11).

[Zhou, 2008] D.-X. Zhou. "Derivative reproducing properties for kernel methods in learning theory". In: *Journal of Computational and Applied Mathematics* 220.1 (2008), pp. 456–463 (cit. on p. 150).

[Zhu et al., 2018] W. Zhu, C. Liu, W. Fan, and X. Xie. "DeepLung: Deep 3D Dual Path Nets for Automated Pulmonary Nodule Detection and Classification". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018, pp. 673–681 (cit. on p. 108).